

ENSURING FAIRNESS IN DIFFICULTY AND CONTENT AMONG PARALLEL ASSESSMENTS GENERATED FROM A TEST-ITEM DATABASE

ABSTRACT

This paper presents research and provides a method to ensure that parallel assessments, that are generated from a large test-item database, maintain equitable difficulty and content coverage each time the assessment is presented

James R. Parry, M.Ed., CPT

Ensuring Fairness in Difficulty and Content Among Parallel Assessments Generated from a Test-item Database

James R. Parry, M.Ed., CPT

Owner/Chief Executive Manager

Compass Consultants, LLC, Virginia, United States of America

May 2020

ABSTRACT

This paper presents research and provides a method to ensure that parallel assessments, that are generated from a large test-item database, maintain equitable difficulty and content coverage each time the assessment is presented. To maintain fairness and validity it is important that all instances of an assessment, that is intended to test the same subject content, are presented to each test-taker without bias. The method described demonstrates how each test-item in an item bank¹ (i) is assigned a difficulty rating using a recognized test-centered² legally defensible cut-score³ determination method, (ii) assigned to either hard, moderate, or easy categories, and (iii) then selected in a stratified random selection⁴ fashion, by content and difficulty, to ensure equal content coverage and difficulty level within an acceptable range of the calculated cut score.

¹ An item bank is an organized collection of items.

² Test-centered methods rely on judgments about test-items, whereas the examinee-centered methods rely on judgments about examinees.

³ To be legally defensible and meet the Standards for Educational and Psychological Testing, a cut score cannot be arbitrarily determined, it must be empirically justified.

⁴ Stratified sampling divides the population or data into groups or blocks. Random samples or selections are taken from each group or block.

EXECUTIVE SUMMARY

With the proliferation of computer based, online and paper-based assessment/testing platforms the storage of test-items (questions) in a database has become commonplace. Without proper preparation of the database, which includes the initial assignment of key fields within the structure of the database, a danger of unfair assessment/testing exists. This unfairness is created by using random selection of test-items from a database without regard to item difficulty and coverage of content that supports the objectives or competencies being assessed. This unfairness is amplified when randomly selecting items from a database to produce parallel forms⁵ of assessments/tests to be given simultaneously to a group or as make-up or retests.

The purpose of this research study is:

- to present evidence that when test-items are selected at random from a database, the resulting parallel forms of the assessments/tests will NOT:
 - cover content equally with each iteration of a parallel assessment/test
 - be equivalent in regard to difficulty of test-items
- to present evidence that when using stratified random selection of test-items from a database, the resulting parallel forms of the assessments/tests will:
 - cover all content equally and at the same difficulty with each iteration of a parallel assessment/test
 - maintain an overall assessment/test difficulty within an acceptable range of the calculated cut score

The paper provides a brief background on test development and procedures for establishing defensible cut or passing scores. In order to conduct the study, three experiments were conducted using both hypothetical and real client test-item difficulty data. The test-item data was entered into a spreadsheet tool developed by James R. Parry that:

- calculates item difficulty based upon the results of a cut-score rating session and assigns a rating of hard, moderate, or easy,
- calculates a cut or passing score for the entire database, and
- provides a final assessment/test design criterion that will maintain equality in both content coverage and difficulty for parallel test forms.

The overall findings conclude that pure random selection of test items will not produce fair parallel forms of assessments/tests but stratified random selection will.

⁵ Parallel forms are different versions of a test that measure the same objectives and yield similar results. (Shrock & Coscarelli, 2007)

Recommendations of the study are:

In order to maintain fairness and ensure parallel forms of assessments/tests are valid, reliable, without bias and defensible when generated from a test-item database:

- Test-items must be constructed using universally recognized standards
- Cut scores should be established using a recognized test-centered method or, if appropriate, a test-taker centered method, because arbitrary methods are not defensible
- Each item in a test-item database should be evaluated by a panel of expert judges and a difficulty score or rating established based upon the agreed upon MAC level of the target test-taker
- Test items should be selected using stratified randomization based upon both topic coverage as well as item difficulty to ensure equitable parallel assessments are generated
- Tests should not be generated in a pure random fashion from a test-item database without regard to content because content coverage will be erratic
- Tests should not be generated in a pure random fashion from test-item database without regard to difficulty of test-items because difficulty among tests will be erratic

TABLE OF CONTENTS

ABSTRACT.....	1
EXECUTIVE SUMMARY	2
INTRODUCTION	5
RESEARCH QUESTIONS.....	7
THE TESTING PROCESS.....	7
TEST DESIGN.....	8
ESTABLISHING A DEFENSIBLE CUT SCORE OR DIFFICULTY RATING	11
Angoff/Modified Angoff Method.....	12
ITEM DATABASES	15
EXPERIMENTAL PROCEDURES.....	16
Basic description of the Questionmark OnDemand assessment platform:.....	16
Design philosophy of the Compass Consultants, LLC spreadsheet tool:	16
Experiment #1A – Random Selection – Hypothetical Data.....	17
Experiment #1B – Stratified Randomization – Hypothetical Data.....	21
Experiment #2A – Random Selection – Real Client #1 Data	25
Experiment #2B – Stratified Randomization – Real Client #1 Data	30
Experiment #3A – Random Selection – Real Client #2 Data	33
Experiment #3B – Stratified Randomization – Real Client #2 Data	37
CONCLUSIONS.....	40
RECOMMENDATIONS	43
LIST OF FIGURES.....	44
LIST OF TABLES.....	45
APPENDIX A.....	A-1
Description of the spreadsheet tool:	A-1
Design philosophy of the Compass Consultants, LLC spreadsheet tool:	A-2
Function of the Spreadsheet Tool.....	A-2
LIST OF FIGURES IN APPENDIX A.....	A-8
LIST OF TABLES IN APPENDIX A.....	A-8
REFERENCES.....	R-1
ACKNOWLEDGEMENTS.....	Ack-1
ABOUT THE AUTHOR.....	Ack-1

INTRODUCTION

Assessments are important evaluation tools used in educational, industrial, government, medical, military and other organizations throughout the world. With the proliferation of electronic/computer databases used to store test-items and generate randomized assessments, comes the inordinate possibility of unfairness. When assessments are intended to be parallel, this unfairness can be either in (i) difficulty, i.e. some assessments are more difficult than others; or (ii) content, i.e. selecting more or less items from one or more topics for every instance of the assessment. If parallel assessments are not equal in both difficulty and content, both the test-taker and the testing organization are at risk. An unfair assessment that consists of mostly “easy” test-items may serve to indicate that a minimally competent candidate is qualified when in fact, they just “got lucky”, whereas an assessment consisting of mostly “hard” test-items, on the same topic or topics as the easy assessment, may deny a candidate, who is minimally qualified or competent, a position or promotion that they truly deserve. Alternatively, what if, because of random item selection, all instances of an assessment do not test the same content? This would provide incorrect assumptions that all candidates “knew” all of the required objectives or competencies when, in fact, they were not tested on all requirements.

Example:

A hypothetical 20-item end of unit assessment on electrical safety is intended to test four topics of equal importance:

- a) Grounding
- b) Lock-out-tag-out
- c) Personal safety equipment
- d) Insulation

Because all topics are considered to be of equal importance, we assume that five test-items should be presented for each topic. The test-item data base contains 100 test-items pertaining to electrical safety, not sorted by topic or by difficulty. The assessment is administered via computer with each student receiving a randomly generated version with a selection criterion of “*select 20 questions at random from topic ‘electrical safety’*”.

Think about all of the possible outcomes of this item selection criterion in terms of both test difficulty and content and the problems associated with this method. With the possibility of generating an unfair assessment, why would test administrators even want to generate a random assessment instead of relying on a single fixed form⁶?

⁶ A fixed-form assessment asks all test-takers to respond to the same questions or tasks in the same order.

There are several advantages for using randomization to select test-items from a database instead of a fixed form which include:

- Anti-cheating – all test-takers do not receive the same test-items so looking at the computer screen or paper of a nearby test-taker will not present an advantage
- Ability to add or retire⁷ test-items – new test-items can be added to or retired from the item database as references or competencies change
- Provides the ability to administer the assessment to the same test-taker multiple times with different test-items
- Allows test administrators to deliver the assessment at different times without fear of pre-test communications among test-takers (i.e. test-taker on East coast of the U.S. calling test-taker on West coast of U.S. about the test contents)

An alternative to pure random selection of test items to generate parallel forms is to use a fixed form but randomize both the test-items and alternatives within the assessment/test. This will produce parallel forms each time and maintain both content and difficulty fairness. Disadvantages to this method are possible overexposure of items to test takers and the inability to produce make-up or retests that would present alternative test items. Because the stem of the test-items remains unchanged there is still a danger of pre-test compromise due to communication among test-takers.

So, what are the dangers of using pure random selection to generate assessments/tests?

- Randomization may produce measurement error⁸ in that some test takers may be presented with a difficult assessment/test and others may receive an easy one. When an assessment/test is used to classify test-takers into groups, two kinds of wrong decisions can occur (Livingston & Zieky, 1982):
 - A test-taker who actually belongs in the lower group can get a score above the passing score
 - A test-taker who actually belongs in the higher group can get a score below the passing score

Classifying someone into the wrong group could lead to less than qualified individuals being promoted or certified or, those who are qualified being denied advancement or certification. These situations could lead to legal challenges that may be difficult to defend.

The item selection method that I propose to alleviate possible unfairness when using randomization to select test-items is described in detail in this paper.

⁷ Once a test-item has been used on an actual assessment it should never be deleted entirely from an item database due to the possibility of future legal proceedings. Retiring the item (if available in the software) retains the item and all statistics in their original form.

⁸ Measurement error in education generally refers to either (1) the difference between what a test score indicates and a student's actual knowledge and abilities or (2) errors that are introduced when collecting and calculating data-based reports, figures, and statistics related to schools and students. (Great Schools Partnership, 2013)

RESEARCH QUESTIONS

1. When test-items which have been assigned a difficulty rating using a recognized test-centered, empirically justified procedure and placed in appropriate topic classification areas at various difficulty levels (easy, moderate, or hard) are randomly selected from a computer based test-item bank, will every iteration of the test generated be presented at a similar difficulty level and cover all topics within objectives adequately?
2. When test-items which have been assigned a difficulty rating using a recognized test-centered, empirically justified procedure and placed in appropriate topic classification areas at various difficulty levels (easy, moderate, or hard) are selected in a stratified-random manner from a computer based test-item bank, will every iteration of the test generated be presented at a similar difficulty level and cover all topics within objectives adequately?

Before any item selection process is attempted it is important to design both effective test-items and test instruments. I will begin with a brief overview of the testing process and design that is the basis for the stratified random selection method that I propose.

THE TESTING PROCESS

I begin with a quote from Steven M. Downing, University of Illinois at Chicago:

“Effective test development requires a systematic, well-organized approach to ensure sufficient validity evidence to support the proposed inferences from test scores.”

(Downing, 2006)

The testing process has many stakeholders which may include members of management, administrators, facilitators, instructors, teachers, test administrators, etc. but the most important stakeholder is the test-taker. If the test is unfair or biased, the test-taker is placed at a disadvantage.

This leads to another quote:

“All tests should be well developed and testing practices, beneficial. There is extensive evidence documenting the effectiveness of well-constructed tests in relation to supporting the validity of the test. The proper use of tests can result in making wiser decisions about individuals and programs than those made without using tests. The improper use of tests, however, can cause considerable harm to test-takers and others affected by test-based decisions.”

(AERA; APA; NCME, 2014)

Whether an assessment/test is norm-referenced⁹ or criterion-referenced¹⁰, it must be fair to all test-takers. If a test is being administered to individual or multiple participants it must be developed and constructed to ensure it is both valid – measures what it is supposed to measure, and reliable - effectively measures anything at all.

TEST DESIGN

The first, and most important step in creating fair, defensible assessments or tests is to ensure validity. All test-items must match the required job skills or certification requirements. Referring to the *Designing Criterion Referenced Tests* flow chart (Figure 1) as a guidance tool, it should be noted that test-item design comes right after the job analysis phase, before course material is created.

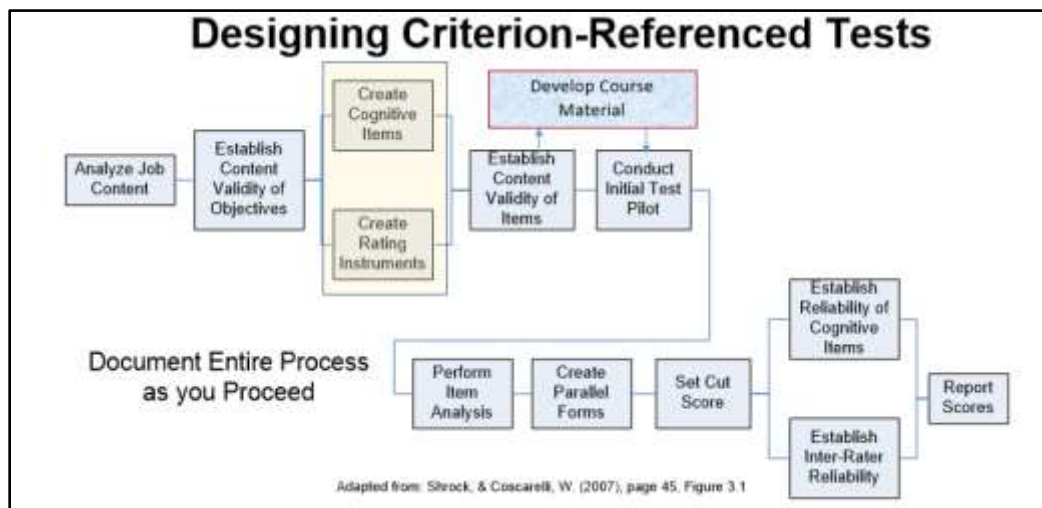


Figure 1 – Designing Criterion-Referenced Tests

The next big step is to determine how many items to develop. To solve this dilemma, ask the following questions:

- How critical are decisions based upon the results of the test?
- What resources (time, money, and personnel) are available for testing?
- How big is the overall objective that is being tested?
- How closely related are the objectives that are being tested?

Table 1, which was adapted from *Criterion-Referenced Test Development, Technical and Legal Guidelines for Corporate Training* (Shrock & Coscarelli, 2007), provides a starting point to help determine the number of test-items required per objective being tested. Once the number of items

⁹ A norm-referenced test compares people in relation to the test performance of one another. It is generally composed of items that will separate the scores of test-takers from one another and is typically used to rank-order select top performers.

¹⁰ A criterion-referenced test compares people to a standard. It is composed of items based on specific objectives or competencies. It defines the performance of each test-taker without regard to the performance of others. It is a test of mastery.

that are required to adequately test an objective has been determined, based on the table, I suggest that number be increased three to five times to allow for an adequate number of items to be available for stratified randomized selection from a test-item database. This ensures that there are an adequate number of items available in each topic to generate several iterations of a parallel assessment without repeating the same items.

Number of Test Items per Objective				
If	—	And	And	Then
The performance objective is:	Critical to safety, life, limb, legal requirements, etc.	From a large objective domain	Unrelated	10-20
			Related	10
		From a small objective domain	Unrelated	5-10
			Related	5
	NOT critical to safety, life, limb, legal requirements, etc.	From a large objective domain	Unrelated	6
			Related	4
		From a small objective domain	Unrelated	2
			Related	1

Adapted from Shrock, S. & Coscarelli, W. 2007

Table 1 - Number of Test-Items per Objective

Typically, there are several topics within an objective, some that are considered more important or critical than other topics. Table 1 can be used as a guide to assist in this decision process. An example may be derived from the hypothetical 20-question end of unit assessment on electrical safety presented in the introduction to this paper. This assessment is designed to test four topics of equal importance:

- a) Grounding
- b) Lock-out-tag-out
- c) Personal safety equipment
- d) Insulation

Because the topics are considered to be of equal importance, 25% of the test-item database is dedicated to each of the topics. To allow for stratified randomization to generate a 20-item assessment, the number of items available in the database for each topic would be as follow:

- a) Grounding 15 - 25
- b) Lock-out-tag-out 15 - 25
- c) Personal safety equipment 15 - 25
- d) Insulation 15 - 25

This requires a test-item database size of between 60 to 100 items to select from to provide five items for each topic on each assessment and ensure different but equal items on each randomly generated assessment.

If all topics are not considered to be equally important or critical then the number of items available per topic, within each objective, would be adjusted, for example if *Grounding* is considered the most important topic it may be ranked as 50% of the assessment with *Lock-out-tag-out* ranked second at 30% and both *Personal safety equipment* and *Insulation* ranked equally important at 10% each. This would require each assessment to be planned and generated as illustrated in table 2.

Topic	% of Test	Number on 20 Question Test	Number Available in Database
Grounding	50	10	30 - 50
Lock-out-tag-out	30	6	18 - 30
Personal safety equipment	10	2	6 - 10
Insulation	10	2	6 - 10

Table 2 - Topic Weights for Electrical Safety Assessment

As can be derived from table 2, it is easy to see how test-item databases can become quite large to ensure randomization. Seems pretty straight forward at this point but what about difficulty? It is important to try to maintain a good balance of easy, moderate and hard items to select from, and include on each test or assessment. When the test-item authors, in consultation with subject matter experts (SMEs), are building the items, they generally will have an idea as to the difficulty of each item with more complex items being more difficult than simple items in most cases. The actual difficulty will not be confirmed until a cut-score rating session is convened.

Another consideration is test length vs. time available. What if, using table 1 as a guide, the test length is determined to be 100 questions. How much time is available for testing? Research and best practice indicated that typical response time for each test-item, on a written test, without references provided, ranges from 42 seconds for an easy item to 90 seconds for a very difficult item. Table 3, based on information in a paper by Phil Higgins (Higgins, 2009), provides a tool to assist in determining time to be allotted for administration of an assessment, not including administrative time (attendance, paperwork, reading test procedures/rules, etc.). Referring to table 3, a typical 100 item 4-alternative multiple-choice assessment that is considered to be of moderate difficulty would require approximately 5,500 seconds or about 1 hour 32 minutes hours to complete, plus administrative time for a total estimated time of about 2 hours. The table is useful as an initial planning tool. Using stratified randomization that defines the actual number of easy, moderate, and hard items selected for each iteration of the assessment, could further refine initial estimates of test time allotted (e.g. 50 item assessment consisting of 20 easy, 20 moderate, and 10 hard test items would yield an approximate test time of 43 minutes). The time it takes for individual test-taker to respond to each item is affected by several factors including, but not limited to language comprehension, education level, reading ability, previous exposure to test item, test-taker motivation, test-taker fatigue, unfamiliarity with test administration system, test anxiety, disabilities, etc. A possible problem with rule of thumb time estimates is that the legality may be challenged unless there is reliable data to back up the estimate. How can the test designer prove that the time allotted is sufficient? A review of statistical information related to the time of response by item, which is

typically available with online test administration software suites, will provide a more accurate estimate once the assessments have been in use.

Overall Test or Item Difficulty	Time Allotted per Test Item		
	Easy	Moderate	Hard
Time per test item	43 Seconds	55 Seconds	61 Seconds
50 Item Test Time	2150 seconds (approx. 36 minutes)	2750 seconds (approx. 46 minutes)	3050 seconds (approx. 51 minutes)
100 Item Test Time	4300 seconds (approx. 72 minutes)	5500 seconds (approx. 92 minutes)	6100 seconds (approx. 102 minutes)

Table 3 - Test Time Tool

It is interesting to note that the U.S. Coast Guard found that if references are allowed to be used during the test, the average time spent per item increases by about 36% (United States Coast Guard, 2015). If the time required is excessive, consideration must be given to omit some topics or rethink the purpose of the test outcome –

- Are the objectives being tested covering too large of a content domain?
- Are all of the topics being tested actually required and supported by the objectives or certification requirements?
- Are the “topics” too far ‘down in the weeds’? In other words, is it a step of a topic at a higher level? (e.g. Unscrewing a light bulb is a step in a larger topic of changing a light bulb).

ESTABLISHING A DEFENSIBLE CUT SCORE OR DIFFICULTY RATING

Establishing a cut score on an assessment or test is also known as standard setting. The ‘standard’ may be in the form of pass/fail, issue/non-issue of a license or certification, award/withhold a credential, etc. The cut score, in order to be legally defensible, cannot be established arbitrarily, it must be empirically justified¹¹ (AERA; APA; NCME, 2014). There are several recognized methods, both test centered and test-taker centered¹², to establish defensible cut scores. Gregory J. Cizek (Cizek, 2006) has done a great job describing and summarizing many of these in his paper *Standard Setting*. One thing to keep in mind, according to the *Standards for Educational and Psychological Testing*, “There can be no single method for determining cut scores for all tests or for all purposes, nor can there be any single set of procedures for establishing their defensibility” (AERA; APA; NCME, 2014) pg. 100.

Most, if not all, of the recognized methods for establishing a cut score are designed around fixed form tests or assessments where the test-items are selected manually and a cut score established on the contents of a single test. If a second, third or subsequent test is generated the cut score

¹¹ Something empirically justified can be provable or verifiable by experience or experiment (Dictionary.com Unabridged, 2020)

¹² Test-taker centered methods require judges to make decisions based on their knowledge of the examinees and their performance

would typically vary because each iteration of the test would be judged individually. The method I propose is a variation of the Modified Angoff Method that evaluates the entire database as a whole, establishes a cut score and difficulty rating for all items within the database and produces a recommended stratified randomized test design that maintains both the cut score as well as covers all topics equally and at similar difficulty with each iteration of a test. The Angoff/Modified Angoff Method with my variations is described as follows:

Angoff/Modified Angoff Method

The Angoff method of determining and setting a cut score for an assessment uses subject matter experts (SMEs) as judges to review each item and assign a score or weight to the item based on the judge's conjecture that a minimally competent performer or a test-taker who is at the minimum acceptable competence (MAC) level required for the job or certification would answer the item correctly. It should be noted that the Angoff Method is sensitive to both difficulty and importance, thus, for a welder, putting on safety glasses is easy, but all would be expected to be able to perform this task – while landing a plane in wind shear is difficult, but all pilots would also be expected to pass this task (Coscarelli, Barrett, Kleeman, & Shrock, 2005). These scores, based on 100 test-takers, are then summed and averaged to assign a difficulty rating to each item. i.e. Six judges score an item as .65, .70, .65, .75, .60, .70. The average of the scores is .675 or 67.5% which translates to an estimation that 67.5% of test-takers at the MAC level would respond correctly to the item. The Angoff Method establishes both a *floor* and *ceiling* score for each item. Typically, the floor score (the lowest a judge is permitted score an item) is based on the number of alternatives for the test-taker to choose from and the ceiling is based upon the results of the judges, who are considered to be experts, average test scores. The ceiling cannot be higher than the judges average scores because a test-taker at the MAC level could not be expected to respond correctly if an expert cannot. The Angoff Method, as with most test-centered methods, is designed to rate the test-items on a single form of a test, so, in theory, each form of the same assessment, all testing the same content, could have a different cut score because each item has a unique Angoff weight (Coscarelli, Barrett, Kleeman, & Shrock, 2005) (e.g. test form A has a cut score of 82%, test form B has a cut score of 80%, and test form C has a cut score of 85%). Each of the cut scores would probably be legally defensible because a recognized standard setting method was used to set the scores but the testing organization may have difficulty explaining the score difference both internally and externally.

In the paper *The Problem of the Saltatory Cut-Score: Some Issues and Recommendations for Applying the Angoff to Test-item Banks* (Coscarelli, Barrett, Kleeman, & Shrock, 2005) the item database was divided into two groups – the first containing all items with an Angoff score less than or equal to the median score and the second group containing all items with scores greater than the median score. The paper explains how item selections were made using simple randomization as well as stratified randomization to generate a fair test. The stratified randomization described, alternates item selection from those items equal to the median Angoff score, those less than or equal to the Angoff score and those greater than the median score. This method guarantees an assessment that is neither extremely difficult or extremely easy. Their conclusions were:

- For low stakes tests randomly sample within the item bank

- For medium stakes tests, one can probably sample within the bank if the distribution is statistically normal, but stratification is safer
- For high stakes tests, one should consider stratification of the sample for increased precision

The methods and results described in the study indicate that random sampling/selection generally will produce tests with the approximate same difficulty when drawn from a small sample but as the size of the sample (database) increases, some sort of stratified selection should be used. The authors warn that as the size of the database increases, so does the chance for errors in selection. Additionally, they recommend stratified random selection as the criticality of the assessment increases. The stratified random selection method that I propose provides greater accuracy in both maintaining the established cut score, presenting test-items at the same difficulty level and selects the same number of items from multiple topics within the item database for each iteration of the parallel assessment. The authors warn, and I concur, the algorithm (or any algorithm) used to select items is only as good as the quality of the Angoff weights (scores) that have been assigned to each item.

My method, which I will call the *Parry Method* for simplicity, is based on the Angoff/Modified Angoff Methods with the following differences (**Note:** There are several variances or adaptations of the Modified Angoff Method in use throughout the testing community):

- The Angoff Method allows any score between 0 and 1 (0% and 100%)
 - Parry Method sets the floor score at the chance guess probability for the number of plausible alternatives available in a multiple-choice (MC) style item (i.e. 4 alternative MC item floor score would be .25 (25%), 3 alternative MC item floor score would be .33 (33%), etc.
 - Parry Method sets the ceiling score at .95 (95%) for all items based on the assumption that if all test-takers (100%) who are considered to be at the MAC level of competence would answer the item correctly the item is a ‘wasted’ item and does not provide much value in discrimination
- Angoff Method requires each judge to “take” the assessment as a typical test-taker would to assign a ceiling score
 - Parry Method eliminates the requirement for the judges to “take” the assessment and sets the ceiling score for each item at .95 (95%). This is probably the most difficult task for the judges because they must put themselves in the mindset and knowledge level of the test-taker at whatever level they have agree upon as the MAC
 - The reason for not having each judge “take” the test is because of the time required to answer all items in large databases and the fact that, if stratified random selection is used to generate each test, there is not a single “test” to take.

- Angoff Method assigns weight or score to each item in a fixed form and sets a cut score based on the average item scores of the panel of judges
 - Parry Method assigns a weight or score to each item in the test-item database. (i.e. if the database has 300 items, each item is evaluated individually without regard to what the final test form will look like) and establishes a cut score for the entire database as a whole
- Angoff Method generally allows the judges to rate the items at any score between 0 and 1 (0% - 100%)
 - Parry Method only allows the judges to rate the items at fixed intervals beginning with the floor score and ending at the .95 ceiling at intervals of .05 (.25, .30, .35, .40, .45, .50, .55, .60, .65, .70, .75, .80, .85, .90, .95)
- Angoff Method requires judges to disregard any alternative/distractor that even someone at the MAC level would not choose before deciding their score
 - Parry method assumes that all stems¹³ and alternatives were designed using recognized item development and review techniques so there should be no “wasted” distractors but requires the judges to comment on the plausibility¹⁴ of existing distractors and recommend changes if necessary
- Angoff Method allows the judges to come together after their initial individual scoring sessions to discuss differences in their ratings and come to a consensus
 - Parry method calculates the standard deviation (SD)¹⁵ among judges scores for each test-item and requires the judges to discuss their ratings as a group if the SD is 10 or greater to attempt to come to a consensus to bring the SD below 10 if possible. If the judges cannot come to a consensus the item is either retired or the judge(s) with the outlier score(s) is/are eliminated from the calculation.

¹³ The stem of a test item presents a single definite and explicit question or problem statement

¹⁴ A plausible alternative is one that has the appearance of being credible or believable, not frivolous.

¹⁵ The standard deviation is a measure of how spread out numbers are from each other.

ITEM DATABASES

With the proliferation of automated (computer) test-item databases and test generation software, many tests are most likely generated randomly, sometimes without regard to difficulty or coverage of required objectives. The most important step in designing test-item databases is the initial topic structure which must allow test-items to be stored with separation of topics, by objective or competency, and further down as necessary. This provides a mechanism for both stratified randomized selection of test-items as well as future psychometric analysis. A typical topic structure is illustrated below:

REPOSITORY NAME

OBJECTIVE 1.0

TOPIC 1.1

SUB-TOPIC 1.1.1

Test-item 1.1.1/1

Test-item 1.1.1/2

TOPIC 1.2

SUB-TOPIC 1.2.1

SUB-TOPIC 1.2.2

OBJECTIVE 2.0

TOPIC 2.1

SUB-TOPIC 2.1.1

Test item 2.1.1/1

Test-item 2.1.1/2

SUB-TOPIC 2.1.2

Test-item 2.1.2/1

The sub-topics can be further divided into three difficulty “buckets” to make selection of test-items at various levels of difficulty clear:

SUB-TOPIC 1.1.1

1.1.1 HARD

Test-item 1.1.1/1

1.1.1 MODERATE

Test-item 1.1.1/5

1.1.1 EASY

Test-item 1.1.1/9

An alternative way to identify difficulty levels of individual test-items may be through the use of metatags¹⁶ of Hard, Moderate, Easy if the test-item database software supports this feature.

¹⁶ Assigning a metatag is a way to index items by specific job tasks, knowledge, skills and abilities, difficulty, etc. to allow for more flexible management and selection of test-items within a large database.

EXPERIMENTAL PROCEDURES

All experiments described in this paper were conducted using either hypothetical or real client test-item data entered into a Questionmark® OnDemand® assessment platform (www.questionmark.com) with test design generated by a proprietary spreadsheet tool designed by James R. Parry, Owner/Chief Executive Manager at Compass Consultants, LLC (www.gocompassconsultants.com). The methods described should work on any test development and delivery platform capable of stratified randomization using either metatags or selection by sub-topic.

Basic description of the Questionmark OnDemand assessment platform: Can assess an unlimited number of test-takers, from anywhere in the world. The platform provides a range of assessment formats including ‘drag and drop’, ‘multiple choice’ and many more. Organizations can conduct a range of assessments across different courses and ability ranges. Tests are automatically marked. Results are instantly compiled. Trends and patterns are easy and quick to spot.

Design philosophy of the Compass Consultants, LLC spreadsheet tool: The tool, described fully in Appendix A, is designed to assist in setting a cut score for an assessment based on the results of a test-centered cut-score rating session that uses a panel of judges or experts to evaluate the difficulty of each test item. It can be used to assist in the design of assessments using the correct response P-value¹⁷ scores returned using classical test theory (CTT)¹⁸ or item response theory (IRT)¹⁹. statistics Additionally, it will determine the number of items from each section at each level of difficulty (hard, moderate or easy) as set by the cut-score rating of each item. This assumption is made for 4-choice, multiple choice items with a floor of 25% and a ceiling of 95%. The difficulty is then assigned based on dividing the difference between 25% and 95% by 3 to arrive at the three difficulty levels. The workbook is designed to accommodate up to ten (10) reviewers on the rating panel. The totals required from each section are based upon the numbers of each level of difficulty available in each section as well as the total number of items available. An assumption is made that if there are fewer items available in any particular section(s) than in other section(s), then that section is of less importance or has significantly fewer objectives. As data for each item is entered in each section, the final test design worksheet is updated automatically.

The tool is designed to establish an initial cut score for a new test-item database. I recommend that after an appropriate number of statistical correct response P-value results are reported, after the database is in use, the recommended cut score be revisited and possibly adjusted using another test-centered method based on actual scores, such as the Bookmark Method²⁰.

¹⁷ P-value is the percentage of test-takers who selected each response. Typically, the correct response p-value is referred to as the difficulty index. (Shrock & Coscarelli, 2007)

¹⁸ Classical test theory (CTT), also known as the true score theory, refers to the analysis of test results based on test scores.

¹⁹ Item response theory (IRT) is a statistical way to analyze responses to tests or questionnaires with the goal of improving measurement of validity and reliability


²⁰ Bookmark Method places all test-items in a booklet, ranging from high to low (easiest to hardest) correct response P-value. Judges review each and place a bookmark at the point they feel the MAC will not answer the next harder item correctly. This point, after discussion and averaging all judges selected break point, becomes the cut score.

Experiment #1A – Random Selection – Hypothetical Data

Referring to the example in the introduction section of a test on electrical safety that covers four topics (Grounding, Lock-Out-Tag-Out, Personal Safety Equipment, Insulation), assume all test-items are placed in a single data file called Fairness Research. All four topics are assumed to be of equal importance so the final test design should test all topics equally.

Using the Modified Parry Method of cut score setting, each item was reviewed²¹ and assigned a difficulty rating with data entered into the spreadsheet tool (Table 4 – partial view)

CUT SCORE CALCULATION TOOL														
Course/Certification Name:			COURSE/CERTIFICATION NAME			Test Name:			TEST NAME					
Facilitator Name/Date:			Facilitator Name/Date			Revision 1 Facilitator Name/Date			Date: mm/dd/yyyy					
Enter Topic/TPO/Subject ID:			1.0 Grounding			Revision 2 Facilitator Name/Date								
This spreadsheet tool is the intellectual property of and Copyright ©2020-2023 by Compass Consultants, LLC. Use is limited to the terms of the End User License Agreement (EULA). This copy is limited to: 30 DAY DEMO.														
Test Item QID	Enter # If New, 0 If Retired	Difficulty Metatag	Average Percentage Correct (Angoff Rating)	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name	Expert 6 Name	Expert 7 Name	Expert 8 Name	Expert 9 Name	Expert 10 Name	Standard Deviation
1.0 E1		Hard	77.50	75	80	80	85	75	70					3.24
1.0 E2		Hard	90.00	95	95	95	85	80	90					6.32
1.0 E3		Hard	79.17	80	70	85	70	85	85					7.36
1.0 E4		Hard	80.83	90	90	75	75	75	80					7.36
1.0 E5		Hard	75.83	90	65	70	70	80	80					8.17
1.0 H1		Hard	48.33	45	45	50	50	55	45					4.08
1.0 H2		Hard	29.17	35	25	25	30	35	25					4.92
1.0 H3		Hard	46.67	25	40	50	50	40	40					8.17
1.0 H4		Hard	48.33	50	45	45	50	50	50					2.58
1.0 H5		Hard	33.33	30	35	30	35	40	45					5.81
1.0 M1		Moderate	68.33	75	70	70	65	65	55					4.08
1.0 M2		Moderate	55.00	60	50	60	55	60	45					6.32
1.0 M3		Moderate	68.83	55	55	60	60	65	70					5.81
1.0 M4		Moderate	70.83	80	70	70	70	65	70					4.92
1.0 M5		Moderate	65.00	65	65	65	65	65	65					0.00
1.0 E1		Hard	82.50	90	85	80	85	80	75					3.24
1.0 E2		Hard	95.00	95	95	95	95	95	95					0.00



Topic Cut Score: 58.00

Moderate Difficulty

Approximate Difficulty Rating

25 - 48.3: Hard

48.4 - 71: Moderate

71.8 - 95: Easy

Standard Deviation

A standard deviation of more than 10 will trigger an alert. Discuss the outliers with the judges who set them to determine why. Change as necessary.

20	Easy	In this section	33%
39	Moderate	In this section	33%
20	Hard	In this section	33%
60	TOTAL		100%

Table 4 -Experiment # 1- Cut Score Calculation Tool (partial view) showing hypothetical data

The cut score for the entire database of 60 items was determined to be 58.02%

The spreadsheet tool recommended that 6.67 items at each difficulty level (hard, moderate, and easy) be used to ensure a cut score of approximately 58.02% was maintained. (Table 5)

Topic	Topic Cut Score & Difficulty	Items in Topic	% of Total Items	Available Hard	% From Topic	Available Mod	% From Topic	Available Easy	% From Topic	Total # Needed From Topic	Use Hard (Calculated)	Use Hard (Actual)	Use Mod (Calculated)	Use Mod (Actual)	Use Easy (Calculated)	Use Easy (Actual)	Topic
1.0	58	60	100.00%	20	33%	20	33%	20	33%	20.00	6.67	7	6.67	6	6.67	7	1.0

Table 5 – Experiment #1 - Recommended Test Difficulty Distribution for Hypothetical Data – Random Selection

²¹ For this phase of the experiment all Angoff/Parry Method difficulty data was hypothetical and assigned to force equal numbers of hard, moderate, and difficult items.

Ignoring the recommendation of even distribution to maintain the cut score, the assessment design function of Questionmark OnDemand was instructed to select 20 items at random from the topic Fairness Research which contained all 60 available test-items (Figure 2).

Question selections
20 random question(s) from topic 'FAIRNESS RESEARCH' including subtopics (Avoid previously delivered)

Figure 2 - Questionmark OnDemand Item Selection Criteria for Hypothetical Data

Thirty (n=30) iterations of a 20-question test were generated as illustrated in tables 6, 7 and 8

Experiment #1 - Random Selection of 20 Items from all 4 topics. Desired target difficulty is 58.02 with 5 items from each topic.																			
Attempt 1		Attempt 2		Attempt 3		Attempt 4		Attempt 5		Attempt 6		Attempt 7		Attempt 8		Attempt 9		Attempt 10	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E3	79.17	1.0 E1	77.50	1.0 E1	77.50	1.0 E1	77.50	1.0 E3	79.17	1.0 E4	80.83	1.0 H2	29.17	1.0 E2	90.00	1.0 E2	90.00	1.0 E2	90.00
1.0 E5	75.83	1.0 E5	75.83	1.0 E2	90.00	1.0 E3	75.83	1.0 E5	75.83	1.0 H5	35.83	1.0 M1	68.33	1.0 E3	79.17	1.0 E5	75.17	1.0 E4	80.83
1.0 H1	48.33	1.0 M1	68.33	1.0 H1	48.33	1.0 H4	48.33	1.0 H5	35.83	1.0 M3	60.83	1.0 M3	60.83	1.0 E4	80.83	1.0 E4	80.83	1.0 H3	40.83
1.0 H5	35.83	1.0 M2	55.00	1.0 H3	40.83	1.0 M1	68.33	2.0 E3	72.50	1.0 M4	70.83	1.0 M4	70.83	1.0 H1	48.33	1.0 H2	29.17	1.0 M4	70.83
1.0 M1	68.33	1.0 M3	60.83	1.0 H4	48.33	2.0 E1	82.50	2.0 E4	82.50	2.0 E3	72.50	1.0 M5	65.00	1.0 H2	29.17	1.0 H3	40.83	2.0 E2	95.00
2.0 E2	95.00	2.0 E1	82.50	1.0 M1	68.33	2.0 E3	72.50	2.0 M1	67.50	2.0 E4	82.50	2.0 E2	95.00	1.0 H3	40.83	1.0 H4	48.33	2.0 H3	46.67
2.0 H3	46.67	2.0 H1	25.00	1.0 M5	65.00	2.0 H3	46.67	2.0 M4	49.17	2.0 E5	77.50	2.0 E3	72.50	1.0 M3	60.83	1.0 H5	35.83	2.0 M1	67.50
2.0 H4	44.17	2.0 M1	67.50	2.0 E1	82.50	2.0 H4	44.17	3.0 E3	72.50	2.0 H1	25.00	2.0 E4	82.50	1.0 M4	70.83	1.0 M1	68.33	2.0 M3	61.67
2.0 M3	61.67	2.0 M3	61.67	2.0 E4	82.50	2.0 M2	68.33	3.0 H5	25.83	2.0 H5	30.00	2.0 H4	44.17	1.0 M5	65.00	1.0 M3	60.83	3.0 E1	80.83
2.0 M2	66.33	2.0 M4	49.17	2.0 H1	25.00	2.0 M5	65.00	3.0 M1	65.00	2.0 M5	65.00	2.0 H5	30.00	2.0 E3	72.50	1.0 M5	65.00	3.0 E2	84.17
2.0 M5	65.00	3.0 E2	84.17	2.0 H3	46.67	3.0 E4	82.50	3.0 M2	49.17	3.0 E1	80.83	2.0 M4	49.17	2.0 H5	30.00	2.0 E5	77.50	3.0 E4	82.50
3.0 E1	80.83	3.0 E3	72.50	2.0 H4	44.17	3.0 E5	72.50	3.0 M3	53.33	3.0 E4	82.50	2.0 M5	65.00	2.0 M2	68.33	2.0 H1	25.00	3.0 H4	33.33
3.0 E2	84.17	3.0 H5	25.83	2.0 H5	30.00	3.0 H4	33.33	3.0 M4	49.17	3.0 E5	72.50	3.0 E1	80.83	2.0 M3	61.67	2.0 H4	44.17	3.0 H5	25.83
3.0 H3	28.33	3.0 M1	65.00	3.0 E4	82.50	3.0 M3	53.33	3.0 M5	61.67	3.0 H3	28.33	3.0 E3	72.50	2.0 M5	65.00	2.0 M3	61.67	3.0 M5	61.67
3.0 H5	25.83	4.0 E1	72.50	3.0 E5	72.50	3.0 M4	49.17	4.0 E1	72.50	3.0 H5	25.83	3.0 H4	33.33	3.0 E2	84.17	3.0 E2	84.17	4.0 E3	72.50
3.0 M3	53.33	4.0 E3	72.50	3.0 H4	33.33	4.0 E3	72.50	4.0 E5	94.17	3.0 M1	65.00	3.0 M2	49.17	3.0 H1	26.67	3.0 H2	46.67	4.0 E4	82.50
3.0 M4	49.17	4.0 H2	30.83	4.0 E1	72.50	4.0 E5	94.17	4.0 H1	28.33	3.0 M3	53.33	4.0 E1	72.50	3.0 M1	65.00	4.0 E4	82.50	4.0 H2	30.83
4.0 E4	82.50	4.0 H3	45.00	4.0 E4	82.50	4.0 H1	28.33	4.0 H3	45.00	3.0 M4	49.17	4.0 H5	45.00	4.0 E2	85.00	4.0 E5	94.17	4.0 M1	50.00
4.0 H2	30.83	4.0 H4	31.67	4.0 H3	45.00	4.0 M4	49.17	4.0 M1	50.00	4.0 E3	72.50	4.0 H5	25.83	4.0 H2	30.83	4.0 H5	25.83	4.0 M2	50.00
4.0 M2	50.00	4.0 M5	54.17	4.0 M4	49.17	4.0 M5	54.17	4.0 M2	50.00	4.0 M1	50.00	4.0 M2	50.00	4.0 M5	54.17	4.0 M4	49.17	4.0 M6	49.17
Difficulty	58.67	Difficulty	58.88	Difficulty	59.33	Difficulty	61.92	Difficulty	58.96	Difficulty	59.04	Difficulty	58.08	Difficulty	60.42	Difficulty	59.54	Difficulty	62.83
Easy	6	Easy	7	Easy	8	Easy	8	Easy	7	Easy	8	Easy	6	Easy	6	Easy	7	Easy	8
Moderate	7	Moderate	8	Moderate	3	Moderate	7	Moderate	9	Moderate	7	Moderate	8	Moderate	8	Moderate	5	Moderate	7
Hard	7	Hard	5	Hard	9	Hard	5	Hard	4	Hard	5	Hard	6	Hard	6	Hard	8	Hard	5
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	3	Topic 1	5	Topic 1	7	Topic 1	4	Topic 1	3	Topic 1	4	Topic 1	5	Topic 1	9	Topic 1	10	Topic 1	4
Topic 2	6	Topic 2	5	Topic 2	6	Topic 2	6	Topic 2	4	Topic 2	7	Topic 2	5	Topic 2	5	Topic 2	4	Topic 2	4
Topic 3	6	Topic 3	4	Topic 3	3	Topic 3	5	Topic 3	7	Topic 3	8	Topic 3	4	Topic 3	3	Topic 3	2	Topic 3	6
Topic 4	3	Topic 4	6	Topic 4	4	Topic 4	5	Topic 4	6	Topic 4	2	Topic 4	4	Topic 4	3	Topic 4	4	Topic 4	6

Table 6 - Experiment #1A - Item Selection - Attempts 1 - 10

Attempt 11		Attempt 12		Attempt 13		Attempt 14		Attempt 15		Attempt 16		Attempt 17		Attempt 18		Attempt 19		Attempt 20	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E1	77.50	1.0 E1	77.50	1.0 E1	77.50	1.0 H2	29.17	1.0 E1	77.50	1.0 E4	80.83	1.0 E2	90.00	1.0 E1	77.50	1.0 E3	79.17	1.0 E2	90.00
1.0 E5	75.83	1.0 E5	75.83	1.0 E5	75.83	1.0 H5	35.83	1.0 E4	80.83	1.0 H3	40.83	1.0 H2	29.17	1.0 E3	79.17	1.0 E5	75.83	1.0 E3	79.17
1.0 H3	40.83	1.0 H5	35.83	1.0 H1	48.33	1.0 M2	55.00	1.0 E5	75.83	1.0 H5	35.83	1.0 H4	48.33	1.0 H1	48.33	1.0 H3	40.83	1.0 E5	75.83
1.0 M1	68.33	1.0 M1	68.33	1.0 H4	48.33	1.0 M5	65.00	1.0 H2	29.17	1.0 M2	55.00	1.0 H5	35.83	1.0 H3	40.83	1.0 M5	65.00	1.0 H1	48.33
1.0 M2	55.00	1.0 M5	65.00	1.0 M2	55.00	2.0 E1	82.50	1.0 H3	40.83	1.0 M5	65.00	1.0 M5	65.00	1.0 M3	60.83	2.0 E1	82.50	1.0 H3	40.83
2.0 E2	95.00	2.0 E5	77.50	1.0 M4	70.83	2.0 E2	95.00	1.0 H4	48.33	2.0 E1	82.50	2.0 E3	72.50	2.0 H1	25.00	2.0 E4	82.50	1.0 M2	55.00
2.0 H1	25.00	2.0 H2	33.33	1.0 M5	65.00	2.0 E3	72.50	1.0 H5	35.83	2.0 E3	72.50	2.0 H2	33.33	2.0 H4	44.17	2.0 E5	77.50	2.0 E4	82.50
2.0 M5	65.00	2.0 H4	44.17	2.0 E3	72.50	2.0 E5	77.50	1.0 M2	55.00	2.0 H4	44.17	2.0 H5	30.00	2.0 H5	30.00	2.0 H3	46.67	2.0 H1	25.00
3.0 E2	84.17	2.0 H5	30.00	2.0 E4	82.50	2.0 H2	33.33	2.0 E5	77.50	2.0 M5	61.67	2.0 M5	65.00	2.0 M1	67.50	2.0 H4	44.17	2.0 H5	30.00
3.0 E5	72.50	3.0 E1	80.83	2.0 E3	77.50	2.0 M1	67.50	2.0 H2	33.33	3.0 E1	80.83	3.0 E2	84.17	3.0 E1	80.83	2.0 H5	30.00	3.0 E2	84.17
3.0 H1	26.67	3.0 E2	84.17	2.0 M2	68.33	3.0 E2	84.17	2.0 H3	46.67	3.0 E4	82.50	3.0 E3	72.50	3.0 E3	72.50	2.0 M2	68.33	3.0 E3	72.50
3.0 H5	28.33	3.0 H5	25.83	2.0 M4	49.17	3.0 E3	72.50	2.0 M4	49.17	3.0 H1	26.67	3.0 H4	33.33	3.0 H4	33.33	2.0 M4	49.17	3.0 E4	82.50
3.0 H4	33.33	3.0 M1	65.00	3.0 E1	80.83	3.0 H2	48.33	3.0 H2	48.33	3.0 H2	48.33	3.0 H5	25.83	3.0 M2	49.17	3.0 E3	72.50	3.0 M1	65.00
4.0 E1	72.50	3.0 M2	49.17	4.0 E1	72.50	3.0 H5	28.33	3.0 M1	65.00	3.0 H4	33.33	3.0 M1	65.00	3.0 M4	49.17	3.0 H3	28.33	3.0 M2	49.17
4.0 E5	94.17	3.0 M4	49.17	4.0 E4	82.50	3.0 H5	25.83	3.0 M3	53.33	3.0 M5	61.67	4.0 E3	85.00	4.0 E5	94.17	3.0 M5	61.67	3.0 M5	61.67
4.0 H2	30.83	4.0 H4	31.67	4.0 E5	94.17	3.0 M2	49.17	3.0 M4	49.17	4.0 E1	72.50	4.0 E5	94.17	4.0 H2	30.83	4.0 E3	72.50	4.0 E1	72.50
4.0 H5	25.83	4.0 M1	50.00	4.0 H3	45.00	3.0 M3	53.33	3.0 M5	67.67	4.0 H1	28.33	4.0 H1	28.33	4.0 H3	45.00	4.0 E4	82.50	4.0 E3	72.50
4.0 M2	50.00	4.0 M2	50.00	4.0 H4	31.67	4.0 E5	94.17	4.0 E3	72.50	4.0 H4	31.67	4.0 H2	30.83	4.0 H5	25.83	4.0 H3	45.00	4.0 E5	94.17
4.0 M3	49.17	4.0 M4	49.17	4.0 M2	50.00	4.0 M2	50.00	4.0 E4	82.50	4.0 M2	50.00	4.0 H5	45.00	4.0 M1	50.00	4.0 H5	25.83	4.0 M1	50.00
4.0 M5	54.17	4.0 M5	54.17	4.0 M5	54.17	4.0 M5	54.17	4.0 H3	45.00	4.0 M5	54.17	4.0 M2	50.00	4.0 M4	49.17	4.0 M1	50.00	4.0 M2	50.00
Difficulty	56.21	Difficulty	54.83	Difficulty	65.08	Difficulty	58.67	Difficulty	56.67	Difficulty	55.42	Difficulty	54.17	Difficulty	52.67	Difficulty	59.00	Difficulty	64.04
Easy	7	Easy	5	Easy	9	Easy	7	Easy	6	Easy	5	Easy	5	Easy	5	Easy	8	Easy	10
Moderate	6	Moderate	9	Moderate	7	Moderate	7	Moderate	6	Moderate	6	Moderate	4	Moderate	6	Moderate	5	Moderate	6
Hard	7	Hard	6	Hard	4	Hard	6	Hard	8	Hard	8	Hard	10	Hard	9	Hard	7	Hard	4
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	3	Topic 1	5	Topic 1	7	Topic 1	4	Topic 1	8	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	4	Topic 1	5
Topic 2	3	Topic 2	4	Topic 2	5	Topic 2	6	Topic 2	4	Topic 2	4	Topic 2	4	Topic 2	4	Topic 2	8	Topic 2	3
Topic 3	5	Topic 3	6	Topic 3	1	Topic 3	7	Topic 3	3	Topic 3	6	Topic 3	5	Topic 3	5	Topic 3	3	Topic 3	6
Topic 4	7	Topic 4	5	Topic 4	7	Topic 4	3	Topic 4	5	Topic 4	5	Topic 4	6	Topic 4	6	Topic 4	5	Topic 4	5

Table 7 - Experiment #1A - Item Selection - Attempts 11 - 20

Attempt 21		Attempt 22		Attempt 23		Attempt 24		Attempt 25		Attempt 26		Attempt 27		Attempt 28		Attempt 29		Attempt 30	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E5	75.83	1.0 E5	75.83	1.0 E3	79.17	1.0 E1	77.50	1.0 E1	77.50	1.0 E1	77.50	1.0 E5	75.83	1.0 E1	77.50	1.0 E2	90.00	1.0 E2	90.00
1.0 H1	48.33	1.0 H1	48.33	1.0 E4	80.83	1.0 E2	90.00	1.0 E4	80.83	1.0 H1	48.33	1.0 H1	48.33	1.0 E3	79.17	1.0 H4	48.33	1.0 H2	29.17
1.0 H4	48.33	1.0 H2	29.17	1.0 H2	29.17	1.0 H5	35.83	1.0 H3	40.83	1.0 H4	48.33	1.0 H2	29.17	1.0 E4	80.83	1.0 M5	65.00	1.0 H3	40.83
1.0 M5	65.00	1.0 H3	40.83	1.0 M4	70.83	2.0 E2	95.00	1.0 H4	48.33	1.0 H5	35.83	1.0 H4	48.33	1.0 H2	29.17	2.0 E3	72.50	1.0 M2	55.00
2.0 E1	82.50	1.0 M1	68.33	1.0 M5	65.00	2.0 E3	72.50	1.0 M3	60.83	1.0 M1	68.33	1.0 M4	70.83	2.0 E1	82.50	2.0 H3	46.67	2.0 E1	82.50
2.0 E2	95.00	1.0 M2	55.00	2.0 E2	95.00	2.0 E4	82.50	2.0 H3	46.67	1.0 M3	60.83	2.0 E3	72.50	2.0 E3	72.50	2.0 H4	44.17	2.0 E4	82.50
2.0 E3	72.50	1.0 M5	65.00	2.0 E3	72.50	2.0 H1	25.00	2.0 H5	30.00	1.0 M5	65.00	2.0 E5	77.50	2.0 E5	77.50	2.0 H5	30.00	2.0 E5	77.50
2.0 E4	82.50	2.0 E4	82.50	2.0 H1	25.00	2.0 H2	33.33	2.0 M4	49.17	2.0 E5	77.50	2.0 H1	25.00	2.0 H1	25.00	2.0 M1	67.50	2.0 H3	46.67
2.0 H4	44.17	2.0 H5	30.00	2.0 H2	33.33	2.0 M2	68.33	3.0 E3	72.50	2.0 M1	67.50	2.0 M2	68.33	2.0 H2	33.33	2.0 M5	65.00	2.0 H5	30.00
2.0 M1	67.50	2.0 M4	49.17	2.0 M5	65.00	2.0 M3	61.67	3.0 E5	72.50	2.0 M2	68.33	3.0 E4	82.50	2.0 H3	46.67	3.0 H2	48.33	2.0 M5	65.00
3.0 E1	80.83	3.0 E2	84.17	3.0 E3	72.50	3.0 E1	80.83	3.0 H1	26.67	2.0 M3	61.67	3.0 H1	26.67	2.0 H4	44.17	3.0 H4	33.33	3.0 E5	72.50
3.0 E4	82.50	3.0 E3	72.50	3.0 E4	82.50	3.0 E3	72.50	3.0 H3	28.33	2.0 M4	49.17	3.0 H2	48.33	2.0 M4	49.17	3.0 M2	49.17	3.0 H2	48.33
3.0 E5	72.50	3.0 E4	82.50	3.0 H1	26.67	3.0 H4	33.33	3.0 H4	33.33	2.0 M5	65.00	3.0 H5	25.83	3.0 E3	72.50	3.0 M4	49.17	3.0 M2	49.17
3.0 H2	48.33	3.0 H2	48.33	3.0 H2	48.33	3.0 M1	65.00	3.0 M5	61.67	3.0 E5	72.50	3.0 M5	53.33	3.0 H1	26.67	3.0 M5	61.67	3.0 M5	53.33
3.0 M4	49.17	3.0 H3	28.33	3.0 M1	65.00	3.0 M4	49.17	4.0 E1	72.50	3.0 H2	48.33	3.0 M4	49.17	4.0 E4	82.50	4.0 E1	72.50	4.0 E2	85.00
3.0 M5	61.67	3.0 M1	65.00	3.0 M2	49.17	4.0 E3	72.50	4.0 E3	72.50	3.0 M4	49.17	4.0 E3	72.50	4.0 E5	94.17	4.0 E5	94.17	4.0 E3	72.50
4.0 E1	72.50	3.0 M3	53.33	3.0 M5	61.67	4.0 H4	31.67	4.0 E5	94.17	4.0 E5	94.17	4.0 E4	82.50	4.0 H1	28.33	4.0 H5	45.00	4.0 E4	82.50
4.0 E4	82.50	4.0 E3	72.50	4.0 E3	72.50	4.0 H5	25.83	4.0 M2	50.00	4.0 H2	30.83	4.0 H3	45.00	4.0 H3	45.00	4.0 H5	25.83	4.0 H4	31.67
4.0 H4	31.67	4.0 H2	30.83	4.0 M4	49.17	4.0 M1	50.00	4.0 M3	49.17	4.0 M2	50.00	4.0 M2	50.00	4.0 M4	49.17	4.0 M2	50.00	4.0 H5	25.83
4.0 M4	49.17	4.0 M5	54.17	4.0 M5	54.17	4.0 M2	50.00	4.0 M5	54.17	4.0 M5	54.17	4.0 M3	49.17	4.0 M5	54.17	4.0 M5	54.17	4.0 M1	50.00
Difficulty	65.63	Difficulty	56.79	Difficulty	59.88	Difficulty	58.62	Difficulty	56.08	Difficulty	58.00	Difficulty	55.04	Difficulty	57.50	Difficulty	55.63	Difficulty	58.50
Easy	10	Easy	6	Easy	7	Easy	8	Easy	7	Easy	3	Easy	6	Easy	9	Easy	4	Easy	8
Moderate	5	Moderate	7	Moderate	8	Moderate	6	Moderate	6	Moderate	12	Moderate	6	Moderate	3	Moderate	8	Moderate	5
Hard	5	Hard	7	Hard	5	Hard	6	Hard	7	Hard	5	Hard	8	Hard	8	Hard	8	Hard	7
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	4	Topic 1	7	Topic 1	5	Topic 1	3	Topic 1	5	Topic 1	7	Topic 1	3	Topic 1	4	Topic 1	3	Topic 1	4
Topic 2	6	Topic 2	3	Topic 2	5	Topic 2	7	Topic 2	3	Topic 2	6	Topic 2	4	Topic 2	8	Topic 2	6	Topic 2	6
Topic 3	6	Topic 3	7	Topic 3	7	Topic 3	5	Topic 3	6	Topic 3	4	Topic 3	6	Topic 3	2	Topic 3	5	Topic 3	4
Topic 4	4	Topic 4	3	Topic 4	3	Topic 4	5	Topic 4	6	Topic 4	3	Topic 4	5	Topic 4	6	Topic 4	6	Topic 4	6

Table 8 - Experiment #1A - Item Selection - Attempts 21 - 30

The item selection difficulty varied widely with each iteration (see ‘Difficulty’ columns in tables 6, 7, and 8). The target cut score/difficulty was 58.02. The randomization produced a difficulty range between 52.67 and 65.63 with average of 58.54. The standard deviation of the scores was 3.10 with a 95% confidence interval²² of 1.10849 which means that the true population mean is between 56.9 and 59.13 of the 30 samples. The kurtosis²³ of the average difficulty is 0.308 and the skewness²⁴ is 0.543. The number of items at each difficulty level from each topic varied with each iteration. Table 9 provides a summary of the statistics for the sample. Figure 3 illustrates the standard distribution curve of the sample.

Sample Difficulty Statistics	
Target Difficulty/Cut Score	58.02
Mean Difficulty	58.54
Median	58.65
Minimum	52.67
Maximum	65.63
Variance Target vs. Mean	0.13
Standard Deviation all Averages	3.10
95% Confidence Score	1.10849085
Kurtosis	0.308023894
Skewness	0.540272895

Table 9 - Difficulty Statistics for Experiment #1A

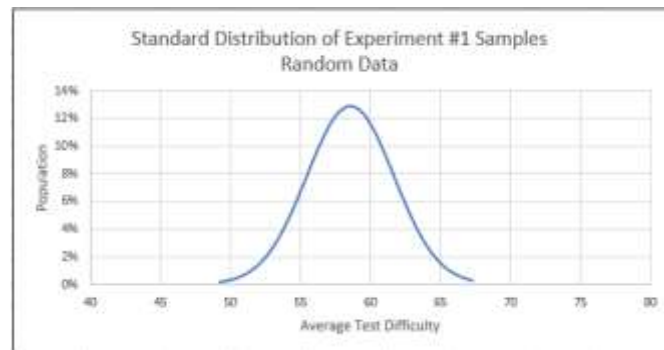


Figure 3 - Standard Distribution of Test Difficulty - Experiment #1A - Random Selection

The content (topic) coverage was erratic as illustrated by the four-color (delineated by sections) display and ‘Total From Topic’ in tables 6, 7, and 8.

Conclusion: All topics were not covered equally in either difficulty or content.

²² A confidence interval is a range of values that you can be fairly sure contains the true mean of the population.

²³ Most often, kurtosis is measured against the normal distribution. If the kurtosis is close to 0, then a normal distribution is often assumed. A low kurtosis indicates a lack of significant outliers. A high kurtosis indicates significant outliers.

²⁴ Skewness is usually described as a measure of a dataset’s symmetry – or lack of symmetry. A perfectly symmetrical data set will have a skewness of 0 which is referred to as “normal” distribution. Negative skew indicates data is skewed left and positive indicates data is skewed right when referring to the “tail”.

Experiment #1B – Stratified Randomization – Hypothetical Data

Using the same item data from experiment #1A, the items were divided into four topics within the main topic and further subdivided into difficulty sub-sub topics based upon the results of the cut score rating tool (Figure 4)

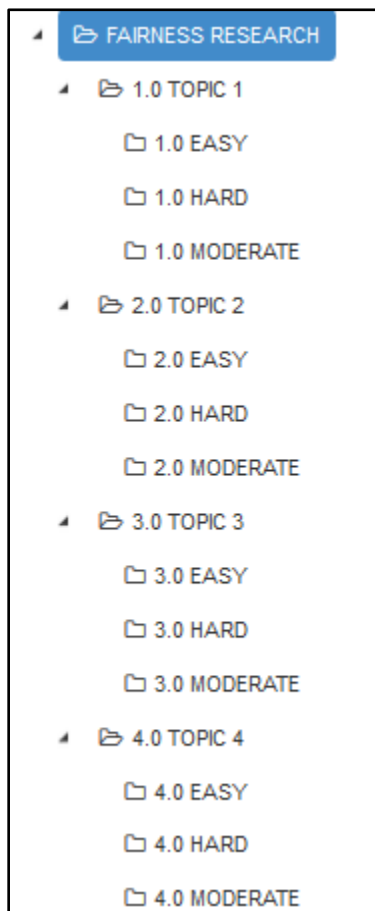


Figure 4 - Topic, Sub-topic Structure Used in Experiments

The final test design recommendations to maintain fairness were as shown in table 10

Topic	Topic Cut Score & Difficulty	Items in Topic	% of Total Items	Available Hard	% From Topic	Available Mod	% From Topic	Available Easy	% From Topic	Total # Needed From Topic	Use Hard [Calculated]	Use Hard [Actual]	Use Mod [Calculated]	Use Mod [Actual]	Use Easy [Calculated]	Use Easy [Actual]	Topic
1.0	62	15	25.00%	5	33%	5	33%	5	33%	5.00	1.67	1	1.67	2	1.67	2	1.0
2	60	15	25.00%	5	33%	5	33%	5	33%	5.00	1.67	2	1.67	1	1.67	2	2
3	56	15	25.00%	5	33%	5	33%	5	33%	5.00	1.67	2	1.67	2	1.67	1	3
4	55	15	25.00%	5	33%	5	33%	5	33%	5.00	1.67	1	1.67	2	1.67	2	4

Table 10 - Recommended Test Design - Stratified Randomization - Experiment #1B

Note: Each topic has a different cut score/difficulty rating but the overall database difficulty remains at 58.02.

Following the recommended test design, the assessment design function of Questionmark OnDemand was instructed to select 20 items in a stratified random fashion from the topic Fairness Research, with five items from each topic, with equal numbers from each level of difficulty (Figure 5).

Question selections
2 random question(s) from topic 'FAIRNESS RESEARCH/1.0 TOPIC 1/1.0 EASY' excluding subtopics
2 random question(s) from topic 'FAIRNESS RESEARCH/1.0 TOPIC 1/1.0 MODERATE' excluding subtopics
1 random question(s) from topic 'FAIRNESS RESEARCH/1.0 TOPIC 1/1.0 HARD' excluding subtopics
2 random question(s) from topic 'FAIRNESS RESEARCH/2.0 TOPIC 2/2.0 EASY' excluding subtopics
2 random question(s) from topic 'FAIRNESS RESEARCH/2.0 TOPIC 2/2.0 HARD' excluding subtopics
1 random question(s) from topic 'FAIRNESS RESEARCH/2.0 TOPIC 2/2.0 MODERATE' excluding subtopics
1 random question(s) from topic 'FAIRNESS RESEARCH/3.0 TOPIC 3/3.0 EASY' excluding subtopics
2 random question(s) from topic 'FAIRNESS RESEARCH/3.0 TOPIC 3/3.0 HARD' excluding subtopics
2 random question(s) from topic 'FAIRNESS RESEARCH/3.0 TOPIC 3/3.0 MODERATE' excluding subtopics
2 random question(s) from topic 'FAIRNESS RESEARCH/4.0 TOPIC 4/4.0 EASY' excluding subtopics
1 random question(s) from topic 'FAIRNESS RESEARCH/4.0 TOPIC 4/4.0 HARD' excluding subtopics
2 random question(s) from topic 'FAIRNESS RESEARCH/4.0 TOPIC 4/4.0 MODERATE' excluding subtopics

Figure 5 - Item Selection Criteria - Experiment #1B

Thirty (n=30) iterations of a 20-question test were generated as illustrated in tables 11, 12, and 13

Experiment #1 - Directed Random Selection of 20 items from all 4 topics. Desired target difficulty is 58.02 with 5 items from each topic.																			
Attempt 1	Attempt 2	Attempt 3	Attempt 4	Attempt 5	Attempt 6	Attempt 7	Attempt 8	Attempt 9	Attempt 10										
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E3	79.17	1.0 E3	79.17	1.0 E1	77.50	1.0 E1	77.50	1.0 E3	79.17	1.0 E2	90.00	1.0 E2	90.00	1.0 E1	77.50	1.0 E2	90.00	1.0 E1	77.50
1.0 E4	80.83	1.0 E4	80.83	1.0 E2	90.00	1.0 E2	90.00	1.0 E5	75.83	1.0 E3	79.17	1.0 E4	80.83	1.0 E2	90.00	1.0 E5	75.83	1.0 E5	75.83
1.0 H5	35.83	1.0 H5	40.83	1.0 H2	29.17	1.0 H2	29.17	1.0 H5	35.83	1.0 H3	40.83	1.0 H2	29.17	1.0 H3	40.83	1.0 H2	29.17	1.0 H1	48.33
1.0 M1	68.33	1.0 M3	60.83	1.0 M3	60.83	1.0 M3	60.83	1.0 M2	55.00	1.0 M1	68.33	1.0 M3	60.83	1.0 M1	68.33	1.0 M4	70.83	1.0 M1	68.33
1.0 M2	55.00	1.0 M4	70.83	1.0 M4	70.83	1.0 M5	65.00	1.0 M5	65.00	1.0 M2	55.00	1.0 M5	65.00	1.0 M2	55.00	1.0 M5	65.00	1.0 M3	60.83
2.0 E2	95.00	2.0 E2	95.00	2.0 E3	72.50	2.0 E1	82.50	2.0 E1	82.50	2.0 E3	72.50	2.0 E3	72.50	2.0 E1	82.50	2.0 E1	82.50	2.0 E3	72.50
2.0 E4	82.50	2.0 E4	82.50	2.0 E5	77.50	2.0 E4	82.50	2.0 E4	82.50	2.0 E4	82.50	2.0 E4	82.50	2.0 E4	82.50	2.0 E5	77.50	2.0 E5	77.50
2.0 H4	44.17	2.0 H4	44.17	2.0 H4	44.17	2.0 H1	25.00	2.0 H1	25.00	2.0 H4	44.17	2.0 H1	25.00	2.0 H1	25.00	2.0 H3	46.67	2.0 H1	25.00
2.0 H5	30.00	2.0 H5	30.00	2.0 H5	30.00	2.0 H2	33.33	2.0 H4	44.17	2.0 H5	30.00	2.0 H3	46.67	2.0 H3	46.67	2.0 H4	44.17	2.0 H4	44.17
2.0 M3	61.67	2.0 M5	65.00	2.0 M5	65.00	2.0 M3	61.67	2.0 M3	61.67	2.0 M5	65.00	2.0 M1	67.50	2.0 M3	61.67	2.0 M2	68.33	2.0 M2	68.33
3.0 E4	82.50	3.0 E1	80.83	3.0 E2	84.17	3.0 E1	80.83	3.0 E5	72.50	3.0 E3	72.50	3.0 E2	84.17	3.0 E4	82.50	3.0 E5	72.50	3.0 E5	72.50
3.0 H1	26.67	3.0 H1	26.67	3.0 H4	33.33	3.0 H1	26.67	3.0 H1	26.67	3.0 H2	48.33	3.0 H1	26.67	3.0 H1	26.67	3.0 H1	26.67	3.0 H3	28.33
3.0 H4	33.33	3.0 H2	48.33	3.0 H5	25.83	3.0 H3	28.33	3.0 H5	25.83	3.0 H4	33.33	3.0 H5	25.83	3.0 H3	28.33	3.0 H5	25.83	3.0 H4	33.33
3.0 M4	49.17	3.0 M1	65.00	3.0 M2	49.17	3.0 M4	49.17	3.0 M1	65.00	3.0 M2	49.17	3.0 M1	65.00	3.0 M1	65.00	3.0 M1	65.00	3.0 M1	65.00
3.0 M5	61.67	3.0 M5	61.67	3.0 M3	53.33	3.0 M5	61.67	3.0 M4	49.17	3.0 M3	53.33	3.0 M5	61.67	3.0 M4	49.17	3.0 M3	53.33	3.0 M3	53.33
4.0 E1	72.50	4.0 E2	85.00	4.0 E4	82.50	4.0 E1	72.50	4.0 E2	85.00	4.0 E1	72.50	4.0 E1	72.50	4.0 E1	72.50	4.0 E2	85.00	4.0 E1	72.50
4.0 E4	82.50	4.0 E4	82.50	4.0 E5	94.17	4.0 E3	72.50	4.0 E5	94.17	4.0 E3	72.50	4.0 E4	82.50	4.0 E2	85.00	4.0 E5	94.17	4.0 E5	94.17
4.0 H1	28.33	4.0 H3	45.00	4.0 H3	45.00	4.0 H4	31.67	4.0 H1	28.33	4.0 H1	28.33	4.0 H5	25.83	4.0 H2	30.83	4.0 H1	28.33	4.0 H2	30.83
4.0 M1	50.00	4.0 M2	50.00	4.0 M1	50.00	4.0 M2	50.00	4.0 M3	49.17	4.0 M1	50.00	4.0 M3	49.17	4.0 M4	49.17	4.0 M1	50.00	4.0 M3	49.17
4.0 M5	49.17	4.0 M5	54.17	4.0 M3	49.17	4.0 M4	49.17	4.0 M1	50.00	4.0 M4	49.17	4.0 M5	54.17	4.0 M5	54.17	4.0 M4	49.17	4.0 M2	50.00
Difficulty	58.42	Difficulty	62.42	Difficulty	59.21	Difficulty	56.50	Difficulty	57.63	Difficulty	57.83	Difficulty	58.38	Difficulty	58.67	Difficulty	60.00	Difficulty	58.37
Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7
Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7
Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6
Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic	Total From Topic
Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5
Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5
Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5
Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5

Table 11 - Experiment #1B Item Selections - Attempts 1 - 10

Attempt 11		Attempt 12		Attempt 13		Attempt 14		Attempt 15		Attempt 16		Attempt 17		Attempt 18		Attempt 19		Attempt 20	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E2	90.00	1.0 E3	75.83	1.0 E5	75.83	1.0 E4	80.83	1.0 E1	77.50	1.0 E4	80.83	1.0 E5	75.83	1.0 E5	75.83	1.0 E4	80.83	1.0 E3	79.17
1.0 E1	77.50	1.0 E1	77.50	1.0 E1	77.50	1.0 E1	77.50	1.0 E3	79.17	1.0 E2	90.00	1.0 E2	90.00	1.0 E3	79.17	1.0 E2	90.00	1.0 E1	77.50
1.0 M3	60.83	1.0 M4	70.83	1.0 M1	88.33	1.0 M3	60.83	1.0 M1	68.33	1.0 M5	60.83	1.0 M1	68.33	1.0 M5	65.00	1.0 M5	65.00	1.0 M4	70.83
1.0 M5	65.00	1.0 M5	65.00	1.0 M5	65.00	1.0 M5	65.00	1.0 M5	65.00	1.0 M2	55.00	1.0 M4	70.83	1.0 M2	55.00	1.0 M5	60.83	1.0 M2	55.00
1.0 H3	40.83	1.0 H5	35.83	1.0 H5	35.83	1.0 H3	40.83	1.0 H5	35.83	1.0 H4	48.33	1.0 H1	48.33	1.0 H5	35.83	1.0 H2	29.17	1.0 H1	48.33
2.0 E4	82.50	2.0 E1	82.50	2.0 E3	72.50	2.0 E1	82.50	2.0 E3	72.50	2.0 E5	77.50	2.0 E3	72.50	2.0 E2	95.00	2.0 E3	72.50	2.0 E2	95.00
2.0 E1	82.50	2.0 E5	77.50	2.0 E5	77.50	2.0 E2	95.00	2.0 E4	82.50	2.0 E1	82.50	2.0 E4	82.50	2.0 E5	77.50	2.0 E2	95.00	2.0 E5	77.50
2.0 H1	25.00	2.0 H4	44.17	2.0 H1	25.00	2.0 H4	44.17	2.0 H1	25.00	2.0 H1	25.00	2.0 H4	44.17	2.0 H5	30.00	2.0 H1	25.00	2.0 H3	46.67
2.0 H4	44.17	2.0 H3	46.67	2.0 H5	30.00	2.0 H1	25.00	2.0 H4	44.17	2.0 H2	33.33	2.0 H2	33.33	2.0 H4	44.17	2.0 H5	30.00	2.0 H2	33.33
2.0 M3	61.67	2.0 M2	68.33	2.0 M3	61.67	2.0 M4	49.17	2.0 M1	67.50	2.0 M5	61.67	2.0 M2	68.33	2.0 M1	67.50	2.0 M5	65.00	2.0 M2	68.33
3.0 E3	72.50	3.0 E1	80.83	3.0 E1	80.83	3.0 E4	82.50	3.0 E5	72.50	3.0 E1	80.83	3.0 E3	72.50	3.0 E2	84.17	3.0 E4	82.50	3.0 E2	84.17
3.0 H3	28.33	3.0 H4	33.33	3.0 H2	48.33	3.0 H1	26.67	3.0 H4	33.33	3.0 H5	28.33	3.0 H3	28.33	3.0 H1	26.67	3.0 H2	48.33	3.0 H5	28.33
3.0 H1	26.67	3.0 H3	28.33	3.0 H4	33.33	3.0 H2	48.33	3.0 H2	48.33	3.0 H5	25.83	3.0 H5	25.83	3.0 H5	25.83	3.0 H1	26.67	3.0 H1	26.67
3.0 M3	53.33	3.0 M3	53.33	3.0 M2	49.17	3.0 M5	61.67	3.0 M1	65.00	3.0 M4	49.17	3.0 M4	49.17	3.0 M3	53.33	3.0 M3	53.33	3.0 M2	49.17
3.0 M4	49.17	3.0 M2	49.17	3.0 M1	65.00	3.0 M2	49.17	3.0 M1	65.00	3.0 M5	61.67	3.0 M1	65.00	3.0 M2	49.17	3.0 M4	49.17	3.0 M1	65.00
4.0 E1	72.50	4.0 E1	72.50	4.0 E5	94.17	4.0 E1	72.50	4.0 E4	82.50	4.0 E4	82.50	4.0 E2	85.00	4.0 E1	72.50	4.0 E2	85.00	4.0 E4	82.50
4.0 E4	82.50	4.0 E5	94.17	4.0 E2	85.00	4.0 E2	85.00	4.0 E1	72.50	4.0 E3	72.50	4.0 E5	94.17	4.0 E5	94.17	4.0 E5	94.17	4.0 E2	85.00
4.0 H1	28.33	4.0 H2	30.83	4.0 H2	30.83	4.0 H5	25.83	4.0 H1	28.33	4.0 H5	25.83	4.0 H2	30.83	4.0 H2	30.83	4.0 H1	28.33	4.0 H4	31.67
4.0 M5	54.17	4.0 M4	49.17	4.0 M5	49.17	4.0 M5	54.17	4.0 M2	50.00	4.0 M2	50.00	4.0 M2	50.00	4.0 M5	54.17	4.0 M3	49.17	4.0 M4	49.17
4.0 M1	50.00	4.0 M5	54.17	4.0 M4	49.17	4.0 M3	49.17	4.0 M5	54.17	4.0 M3	49.17	4.0 M3	49.17	4.0 M3	49.17	4.0 M2	50.00	4.0 M5	54.17
Difficulty	57.38	Difficulty	59.50	Difficulty	58.71	Difficulty	58.79	Difficulty	59.29	Difficulty	57.04	Difficulty	60.21	Difficulty	58.25	Difficulty	59.01	Difficulty	60.25
Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7
Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7
Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5
Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5
Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5
Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5

Table 12 - Experiment #1B Item Selections - Attempts 11 - 20

Attempt 21		Attempt 22		Attempt 23		Attempt 24		Attempt 25		Attempt 26		Attempt 27		Attempt 28		Attempt 29		Attempt 30	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E2	75.83	1.0 E4	80.00	1.0 E3	79.17	1.0 E1	77.50	1.0 E4	80.83	1.0 E1	77.50	1.0 E3	79.17	1.0 E5	75.83	1.0 E3	79.17	1.0 E5	75.83
1.0 E2	90.00	1.0 E3	79.17	1.0 E5	75.83	1.0 E2	90.00	1.0 E1	77.50	1.0 E4	80.83	1.0 E5	75.83	1.0 E5	75.83	1.0 E1	77.50	1.0 E3	79.17
1.0 M4	70.83	1.0 M5	65.00	1.0 M5	65.00	1.0 M5	65.00	1.0 M5	60.83	1.0 M4	70.83	1.0 M3	60.83	1.0 M1	68.33	1.0 M5	65.00	1.0 M4	70.83
1.0 M2	55.00	1.0 M3	60.83	1.0 M3	60.83	1.0 M2	55.00	1.0 M2	55.00	1.0 M2	55.00	1.0 M1	68.33	1.0 M2	55.00	1.0 M2	55.00	1.0 M3	60.83
1.0 H5	48.33	1.0 H2	29.17	1.0 H2	29.17	1.0 H3	40.83	1.0 H5	35.83	1.0 H2	29.17	1.0 H5	35.83	1.0 H3	48.33	1.0 H1	48.33	1.0 H2	29.17
2.0 E4	82.50	2.0 E3	72.50	2.0 E2	95.00	2.0 E5	77.50	2.0 E2	95.00	2.0 E5	77.50	2.0 E1	82.50	2.0 E2	95.00	2.0 E3	72.50	2.0 E5	77.50
2.0 E5	77.50	2.0 E4	82.50	2.0 E3	72.50	2.0 E3	72.50	2.0 E1	82.50	2.0 E4	82.50	2.0 E5	77.50	2.0 E1	82.50	2.0 E5	77.50	2.0 E3	72.50
2.0 H2	33.33	2.0 H1	25.00	2.0 H4	44.17	2.0 H4	44.17	2.0 H3	46.67	2.0 H1	25.00	2.0 H1	25.00	2.0 H3	46.67	2.0 H4	44.17	2.0 H1	25.00
2.0 H5	30.00	2.0 H2	33.33	2.0 H3	46.67	2.0 H2	33.33	2.0 H1	25.00	2.0 H3	46.67	2.0 H4	44.17	2.0 H4	44.17	2.0 H2	33.33	2.0 H2	33.33
2.0 M4	49.17	2.0 M5	65.00	2.0 M5	61.67	2.0 M1	67.50	2.0 M5	65.00	2.0 M4	49.17	2.0 M4	49.17	2.0 M3	61.67	2.0 M5	65.00	2.0 M3	61.67
3.0 E2	84.17	3.0 E3	72.50	3.0 E1	80.83	3.0 E5	72.50	3.0 E4	82.50	3.0 E1	84.17	3.0 E4	82.50	3.0 E2	84.17	3.0 E3	72.50	3.0 E5	72.50
3.0 H4	33.33	3.0 H1	26.67	3.0 H1	26.67	3.0 H3	28.33	3.0 H4	33.33	3.0 H2	48.33	3.0 H4	33.33	3.0 H3	28.33	3.0 H3	28.33	3.0 H2	48.33
3.0 H2	48.33	3.0 H5	28.33	3.0 H2	48.33	3.0 H1	26.67	3.0 H2	48.33	3.0 H1	26.67	3.0 H5	28.33	3.0 H3	28.33	3.0 H2	48.33	3.0 H5	28.33
3.0 M5	61.67	3.0 M5	61.67	3.0 M1	65.00	3.0 M5	61.67	3.0 M5	53.33	3.0 M5	61.67	3.0 M1	65.00	3.0 M2	49.17	3.0 M4	49.17	3.0 M3	53.33
3.0 M4	49.17	3.0 M3	53.33	3.0 M2	49.17	3.0 M4	49.17	3.0 M5	61.67	3.0 M3	53.33	3.0 M5	61.67	3.0 M3	53.33	3.0 M1	65.00	3.0 M1	65.00
4.0 E2	85.00	4.0 E2	85.00	4.0 E4	82.50	4.0 E5	94.17	4.0 E2	85.00	4.0 E1	72.50	4.0 E2	85.00	4.0 E5	94.17	4.0 E1	72.50	4.0 E5	94.17
4.0 E5	94.17	4.0 E5	94.17	4.0 E2	85.00	4.0 E2	85.00	4.0 E5	94.17	4.0 E2	85.00	4.0 E3	72.50	4.0 E4	82.50	4.0 E5	94.17	4.0 E1	72.50
4.0 H5	25.83	4.0 H5	25.83	4.0 H4	31.67	4.0 H2	30.83	4.0 H1	28.33	4.0 H2	30.83	4.0 H3	25.83	4.0 H5	25.83	4.0 H5	25.83	4.0 H3	45.00
4.0 M4	49.17	4.0 M4	49.17	4.0 M1	50.00	4.0 M1	50.00	4.0 M1	50.00	4.0 M2	50.00	4.0 M1	50.00	4.0 M2	50.00	4.0 M2	50.00	4.0 M3	49.17
4.0 M3	49.17	4.0 M5	54.17	4.0 M4	49.17	4.0 M5	54.17	4.0 M2	50.00	4.0 M4	49.17	4.0 M4	49.17	4.0 M5	54.17	4.0 M5	54.17	4.0 M1	50.00
Difficulty	59.63	Difficulty	57.17	Difficulty	59.92	Difficulty	58.79	Difficulty	60.54	Difficulty	57.79	Difficulty	57.83	Difficulty	60.00	Difficulty	59.46	Difficulty	58.21
Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7	Easy	7
Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7	Moderate	7
Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6	Hard	6
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5
Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5	Topic 2	5
Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5
Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5	Topic 4	5

Table 13 - Experiment #1B Item Selections - Attempts 21 - 30

The item selection difficulty remained relatively constant as desired with each iteration (see ‘Difficulty’ columns in tables 11, 12, and 13). The target cut score/difficulty was 58.02. The stratified randomization produced a difficulty range between 56.57 and 62.42 with average of 58.84. The standard deviation of the scores was 1.24 with a 95% confidence interval of 0.4456 which means that the true population mean is between 58.39 and 59.29 of the 30 samples. The kurtosis of the

average difficulty is 0.987 and the skewness is 0.537. The number of items at each difficulty level from each topic varied with each iteration. Table 14 provides a summary of the statistics for the sample. Figure 6 illustrates the standard distribution curve of the sample.

Sample Difficulty Statistics	
Target Cut Score	58.02
Mean difficulty	58.84
Median	58.75
Minimum	56.50
Maximum	62.42
Variance Target vs. Mean	0.34
Standard Deviation all Averages	1.24
95% Confidence Score	0.445089638
Kurtosis	0.986867811
Skewness	0.537266918

Table 14 - Difficulty Statistics for Experiment #1B

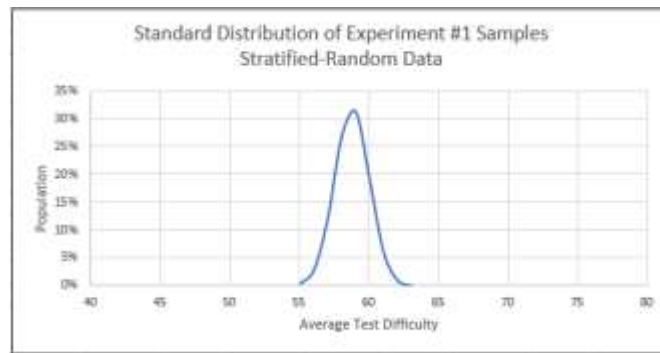


Figure 6 - Standard Distribution of Test Difficulty - Experiment #1B - Stratified Random Selection

The content (topic) coverage was equal as stratified for each iteration as illustrated by the four-color (delineated by sections) display and 'Total From Topic' columns in tables 11, 12, and 13.

Conclusion: All topics were covered equally as desired in both difficulty and content. Comparing the distribution of test difficulty scores in figures 3 and 6 shows that the stratified randomization consistently produced tests well within an acceptable range to meet the desired cut score of 58.02.

Experiment #2A – Random Selection – Real Client #1 Data

Note: Client Test-item QIDs replaced to protect confidentiality

Using cut score (Parry Method) results from real client data (Tables 15, 16, & 17 – partial views), the items were divided into three topics within the main topic and further subdivided into difficulty sub-sub topics based upon the results of the cut score rating tool. The cut score/difficulty for the entire database (71 items) was determined to be 76.13% by averaging all three topic cut scores.

CUT SCORE CALCULATION TOOL

Course/Certification Name:

COURSE/CERTIFICATION NAME

Test Name:

TEST NAME

Facilitator Name/Date:

Facilitator Name/Date

Revision 1 Facilitator Name/Date

Revision 2 Facilitator Name/Date

Date: mm/dd/yyyy

Enter Topic/TPO/Subject ID:

Topic 1

Revision 2 Facilitator Name/Date

This spreadsheet is the intellectual property of and Copyright © 2019 by Compass Consultants, LLC. Use is limited to the terms of the End User License Agreement (EULA). This report is for use by the client only.

Test Item QID	Enter # If New, # If Retired	Difficulty Metatag	Average Percentage Correct (Angoff Rating)	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name	Expert 6 Name	Expert 7 Name	Expert 8 Name	Expert 9 Name	Expert 10 Name	Standard Deviation
1.0 M1		Moderate	63.00	70	65	60	50	70						8.37
1.0 E1			80.00	75	75	80	75	95						9.66
1.0 E2			75.00	75	85	80	50	85						14.58
1.0 M2		Moderate	67.00	70	50	70	70	75						8.75
	R			70	50	70	70	85						
1.0 E3			77.00	75	65	75	95	75						10.95
1.0 M3		Moderate	69.00	70	60	75	70	70						3.48
1.0 E4			76.00	70	70	85	90	85						10.84
1.0 E5			94.00	90	95	95	95	95						2.34
	R			80	50	75	50	60						
1.0 E6			89.00	85	95	90	90	85						4.38
1.0 E7			79.00	80	80	70	80	80						4.47
1.0 M4		Moderate	71.00	75	60	80	60	80						10.25
1.0 E8			91.00	85	95	95	95	85						5.48
1.0 E9			83.00	90	95	80	70	80						9.75
1.0 E10			77.00	75	75	65	90	80						8.68
1.0 E11			78.00	65	85	75	90	75						9.75
1.0 E12			75.00	70	80	70	85	70						9.00

Topic Cut Score

78.80

None

Approximate Difficulty Rating

25 - 48.3 Hard

48.4 - 71 Moderate

71.8 - 95 Easy

Standard Deviation

A standard deviation of more than 10 will trigger an alert. Discuss the outliers with the judges who set them to determine why. Change as necessary.

34	Easy	In this section	78%
4	Moderate	In this section	22%
0	Hard	In this section	0%
38	TOTAL		100%

Table 15 - Experiment #2A - Difficulty Calculations - Real Data - Topic 1

CUT SCORE CALCULATION TOOL

Course/Certification Name:

COURSE/CERTIFICATION NAME

Test Name:

TEST NAME

Facilitator Name/Date:

Facilitator Name/Date

Revision 1 Facilitator Name/Date

Revision 2 Facilitator Name/Date

Date: mm/dd/yyyy

Enter Topic/TPO/Subject ID:

Topic 2

This spreadsheet is the intellectual property of and Copyright © 2019 by Compass Consultants, LLC. Use is restricted to the terms of the End User License Agreement (EULA). This report is for use by the client only.

Test Item QID	Enter # If New, # If Retired	Difficulty Metatag	Average Percentage Correct (Angoff Rating)	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name	Expert 6 Name	Expert 7 Name	Expert 8 Name	Expert 9 Name	Expert 10 Name	Standard Deviation
2.0 E1			85.00	80	80	95	90	70						9.75
2.0 E2			92.00	85	95	95	95	90						4.47
2.0 E3			76.00	70	65	75	90	80						9.62
2.0 E4			75.00	70	85	70	80	70						7.07
2.0 M1		Moderate	63.00	65	75	60	60	55						7.58
2.0 M2		Moderate	48.75		65	50	40	40						11.81
2.0 E5			74.00	70	70	75	85	70						6.52
2.0 E6			80.00	80	75	70	95	80						8.35
2.0 E7			75.00	75	85	65	80	70						7.91
2.0 E8			81.00	85	90	85	80	85						9.82
2.0 E9			89.00	85	95	85	95	85						5.48
2.0 E10			83.00	80	85	80	90	80						4.47
2.0 M3		Moderate	67.00	75	70	60	70	60						6.71
2.0 E11			82.50		80	70	95	85						10.41
2.0 E12			90.00	85	95	90	95	85						5.00
2.0 E13			79.00	75	90	75	75	80						6.53
2.0 E14			90.00	75	95	95	95	90						8.66
2.0 E15			82.00	75	95	75	85	80						6.87

Compass Consultants, LLC

Topic Cut Score

74.00

Moderate Difficulty

Approximate Difficulty Rating

25 - 48.3 Hard

48.4 - 71 Moderate

71.8 - 95 Easy


Standard Deviation

A standard deviation of more than 10 will trigger an alert. Discuss the outliers with the judges who set them to determine why. Change as necessary.

22	Easy	In this section	67%
10	Moderate	In this section	39%
1	Hard	In this section	3%
33	TOTAL		100%

Table 16 - Experiment #2A - Difficulty Calculations - Real Data - Topic 2

CUT SCORE CALCULATION TOOL														
Course/Certification Name:			COURSE/CERTIFICATION NAME			Test Name:		TEST NAME						
Facilitator Name/Date:			Facilitator Name/Date			Revision 1 Facilitator Name/Date				Date: mm/dd/yyyy				
Enter Topic/TPO/Subject ID:			Topic 3			Revision 2 Facilitator Name/Date								
This tool is designed to be used by the user of the tool. It is not intended to be used by the tool. The user of the tool is responsible for the results. The user of the tool is responsible for the results. The user of the tool is responsible for the results.														
Test Item QID	Enter # If New, & If Revised	Difficulty Metatag	Average Percentage Correct (Angoff Rating)	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name	Expert 6 Name	Expert 7 Name	Expert 8 Name	Expert 9 Name	Expert 10 Name	Standard Deviation
3.0.E1			90.00	80	95	90	95	90						8.12
3.0.E2			87.00	75	95	90	90	85						7.58
3.0.E3			84.00	75	85	75	85	80						10.25
3.0.E4			74.00	75	85	60	80	70						8.63
3.0.E5			79.00	80	85	65	90	75						8.63
3.0.E6			73.00	80	75	75	85	70						5.70
3.0.E7			83.00	80	90	70	90	85						8.37
	R			75	70	60	95	80						
3.0.E8			72.00	80	75	65	80	60						8.08
3.0.E9			89.00	80	95	90	95	85						8.52
3.0.E10			85.00	80	85	85	95	80						5.12
3.0.M1		Moderate	57.50		70	50	60	50						9.57
3.0.E11			84.00	80	95	90	70	85						8.63
3.0.E12			85.00	85	95	70	90	85						9.95
3.0.M2		Moderate	61.00	65	70	50	50	70						10.25
3.0.E13			72.00	75	85	70	60	70						8.08
3.0.E14			83.00	80	90	70	95	80						5.72
3.0.E15			73.00	70	80	80	75	80						8.37



Topic Cut Score

77.00

Approximate Difficulty Rating

25 - 48.5 Hard

48.6 - 71 Moderate

71.8 - 95 Easy

Standard Deviation

A standard deviation of more than 10 will trigger an alert. Discuss the outliers with the judges who set them to determine why. Change as necessary.

17 Easy In this section 85%

3 Moderate In this section 15%

0 Hard In this section 0%

20 TOTAL 100%

Table 17 - Experiment #2A - Difficulty Calculations - Real Data - Topic 3

Topic 1 consisted of 18 test-items with a difficulty rating of 78 (Easy). Topic 2 consisted of 33 test-items with a difficulty rating of 74 (Moderate). Topic 3 consisted of 20 test-items with a difficulty rating of 77 (Easy). Each topic had a mix of hard, moderate, and easy items. (Table 18)

Topic	Topic Cut Score & Difficulty	Items in Topic	% of Total Items	Available Hard	% From Topic	Available Mod	% From Topic	Available Easy	% From Topic
Topic 1	78	18	25.35%	0	0%	4	22%	14	78%
Topic 2	74	33	46.48%	1	3%	10	30%	22	67%
Topic 3	77	20	28.17%	0	0%	3	15%	17	85%

Table 18 - Experiment #2 - Item Difficulty Distribution by Topic

Referring to the design philosophy of the Compass Consultants spreadsheet tool as to the number of items drawn from each section, it appears that topic 2 was considered to be the ‘most’ important with 46.48% of the available items, topic 3 was the next ‘most’ important with 28.17% and topic 1 was the least important with 25.35%.

The final test design to maintain fairness in both content and difficulty is shown in table 19.

Total # Needed From Topic	Use Hard (Calculated)	Use Hard (Actual)	Use Mod (Calculated)	Use Mod (Actual)	Use Easy (Calculated)	Use Easy (Actual)	Topic
5.07	0.00	0	1.13	1	3.94	4	Topic 1
9.30	0.28	1	2.82	3	6.20	6	Topic 2
5.63	0.00	0	0.85	1	4.79	4	Topic 3

Table 19 - Experiment #2 - Recommended Test Design

Ignoring the recommendation of item distribution to maintain the cut score, the assessment design function of Questionmark OnDemand was instructed to select 20 items at random from the single topic containing all of the test-items. Tables 20, 21, and 22 present the results.

Experiment #2 - Random Selection of 20 Items from all 3 topics. Real Client Data. Desired target difficulty is 76.13.																			
Attempt 1		Attempt 2		Attempt 3		Attempt 4		Attempt 5		Attempt 6		Attempt 7		Attempt 8		Attempt 9		Attempt 10	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E10	77.00	1.0 E1	80.00	1.0 E10	77.00	1.0 E1	80.00	1.0 E12	75.00	1.0 E10	77.00	1.0 E12	75.00	1.0 E12	75.00	1.0 E1	75.00	1.0 E1	80.00
1.0 E13	76.00	1.0 E10	77.00	1.0 E11	78.00	1.0 E11	78.00	1.0 E2	75.00	1.0 E14	79.00	1.0 E14	79.00	1.0 E13	76.00	1.0 E4	76.00	1.0 E10	77.00
1.0 E3	77.00	1.0 E11	78.00	1.0 E2	75.00	1.0 E7	78.00	1.0 E7	78.00	1.0 E5	94.00	1.0 E9	83.00	1.0 E2	75.00	1.0 E7	78.00	1.0 E14	79.00
1.0 E4	76.00	1.0 E5	94.00	1.0 E6	89.00	1.0 E8	91.00	1.0 E9	83.00	1.0 E6	89.00	1.0 M1	63.00	1.0 E5	77.00	1.0 M3	69.00	1.0 E4	76.00
1.0 E8	91.00	1.0 E9	83.00	2.0 E1	83.00	1.0 M1	63.00	2.0 E10	83.00	1.0 E9	83.00	2.0 E10	83.00	1.0 E7	78.00	2.0 E16	83.00	1.0 M1	63.00
2.0 E1	83.00	1.0 M1	63.00	2.0 E13	79.00	1.0 M4	71.00	2.0 E16	83.00	1.0 M3	69.00	2.0 E11	82.50	2.0 E1	83.00	2.0 E3	92.00	1.0 M5	69.00
2.0 E11	82.50	2.0 E14	90.00	2.0 E2	92.00	2.0 E14	90.00	2.0 E20	80.00	2.0 E1	83.00	2.0 E13	79.00	2.0 E10	83.00	2.0 E3	76.00	2.0 E1	83.00
2.0 E12	90.00	2.0 E15	82.00	2.0 E20	80.00	2.0 E16	83.00	2.0 E4	75.00	2.0 E10	83.00	2.0 E14	90.00	2.0 E13	79.00	2.0 E4	75.00	2.0 E17	79.00
2.0 E15	82.00	2.0 E16	83.00	2.0 E3	76.00	2.0 E19	86.00	2.0 E5	74.00	2.0 E12	90.00	2.0 E17	79.00	2.0 E3	76.00	2.0 E8	81.00	2.0 E18	81.00
2.0 E2	92.00	2.0 E17	79.00	2.0 E4	75.00	2.0 E2	92.00	2.0 E8	81.00	2.0 E17	79.00	2.0 E21	78.00	2.0 E5	74.00	2.0 E9	89.00	2.0 E5	74.00
2.0 E4	75.00	2.0 E21	78.00	2.0 E7	75.00	2.0 E3	76.00	2.0 M10	56.25	2.0 E20	80.00	2.0 E5	74.00	2.0 E8	81.00	2.0 M1	63.00	2.0 E6	80.00
2.0 E5	74.00	2.0 E4	75.00	2.0 E9	80.00	2.0 E4	75.00	2.0 M1	63.00	2.0 E5	74.00	2.0 E6	80.00	2.0 H1	46.25	2.0 M5	67.00	2.0 E8	81.00
2.0 E7	75.00	2.0 E6	80.00	2.0 H1	46.25	2.0 E5	74.00	2.0 M5	67.00	2.0 E8	81.00	2.0 H1	46.25	2.0 M8	52.50	2.0 M5	68.00	2.0 M1	63.00
2.0 E9	89.00	2.0 H1	46.25	2.0 M3	67.00	2.0 E7	75.00	2.0 M6	53.75	2.0 M3	67.00	2.0 M3	67.00	3.0 E1	90.00	3.0 E10	85.00	2.0 M3	67.00
2.0 M9	70.00	2.0 M4	53.00	2.0 M8	52.50	2.0 M4	53.00	2.0 M7	66.25	2.0 M6	53.75	2.0 M4	53.00	3.0 E10	85.00	3.0 E2	87.00	2.0 M4	53.00
3.0 E17	72.00	3.0 M8	52.50	3.0 E10	85.00	2.0 M9	70.00	3.0 E12	85.00	3.0 E15	73.00	2.0 M6	53.75	3.0 E14	83.00	3.0 E3	84.00	2.0 M8	52.50
3.0 E2	87.00	3.0 E12	85.00	3.0 E1	90.00	3.0 E13	72.00	3.0 E14	83.00	3.0 E15	90.00	2.0 M8	52.50	3.0 E16	75.00	3.0 E4	74.00	3.0 E13	72.00
3.0 E9	89.00	3.0 E14	83.00	3.0 E12	85.00	3.0 E14	83.00	3.0 E6	74.00	3.0 E6	75.00	3.0 E12	85.00	3.0 E17	72.00	3.0 E7	83.00	3.0 E15	73.00
3.0 M1	57.50	3.0 E15	73.00	3.0 E7	83.00	3.0 E16	75.00	3.0 E5	79.00	3.0 E8	72.00	3.0 E15	73.00	3.0 E5	79.00	3.0 E8	72.00	3.0 E2	87.00
3.0 M3	57.50	3.0 M3	57.50	3.0 M2	61.00	3.0 E6	73.00	3.0 M3	57.50	3.0 M2	61.00	3.0 E7	83.00	3.0 M1	57.50	3.0 M5	57.50	3.0 E3	84.00
Difficulty	78.63	Difficulty	74.61	Difficulty	76.44	Difficulty	76.90	Difficulty	73.59	Difficulty	77.54	Difficulty	72.95	Difficulty	74.86	Difficulty	76.73	Difficulty	73.68
Easy	17	Easy	15	Easy	16	Easy	16	Easy	14	Easy	16	Easy	14	Easy	17	Easy	15	Easy	14
Moderate	3	Moderate	4	Moderate	3	Moderate	4	Moderate	6	Moderate	4	Moderate	5	Moderate	3	Moderate	5	Moderate	6
Hard	0	Hard	1	Hard	1	Hard	0	Hard	0	Hard	0	Hard	1	Hard	1	Hard	0	Hard	0
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	5	Topic 1	6	Topic 1	4	Topic 1	6	Topic 1	4	Topic 1	6	Topic 1	4	Topic 1	5	Topic 1	4	Topic 1	6
Topic 2	10	Topic 2	10	Topic 2	11	Topic 2	10	Topic 2	11	Topic 2	9	Topic 2	13	Topic 2	8	Topic 2	9	Topic 2	10
Topic 3	5	Topic 3	4	Topic 3	5	Topic 3	4	Topic 3	5	Topic 3	5	Topic 3	3	Topic 3	7	Topic 3	7	Topic 3	4

Table 20 - Experiment #2A - Item Selection - Attempts 1 - 10

Attempt 11		Attempt 12		Attempt 13		Attempt 14		Attempt 15		Attempt 16		Attempt 17		Attempt 18		Attempt 19		Attempt 20	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E14	79.00	1.0 E13	76.00	1.0 E10	77.00	1.0 E10	77.00	1.0 E12	75.00	1.0 E1	80.00	1.0 E4	76.00	1.0 E12	75.00	1.0 E10	77.00	1.0 E10	77.00
1.0 E3	77.00	1.0 E8	91.00	1.0 E14	79.00	1.0 E12	75.00	1.0 E3	77.00	1.0 E2	75.00	1.0 E9	83.00	1.0 E7	78.00	1.0 E12	75.00	1.0 E4	76.00
1.0 E9	83.00	1.0 E9	83.00	1.0 E6	89.00	1.0 E4	76.00	1.0 E5	94.00	1.0 E3	77.00	2.0 E13	79.00	1.0 E8	91.00	1.0 E12	80.00	1.0 E5	94.00
1.0 M1	63.00	1.0 M1	63.00	1.0 E9	83.00	1.0 E8	91.00	1.0 E7	78.00	1.0 M2	67.00	2.0 E14	90.00	2.0 E10	83.00	1.0 E3	77.00	1.0 E6	89.00
1.0 M4	71.00	1.0 M3	69.00	2.0 E11	82.50	1.0 M1	63.00	1.0 E8	91.00	1.0 M4	71.00	2.0 E17	79.00	2.0 E11	82.50	1.0 E5	94.00	1.0 M1	63.00
2.0 E1	83.00	1.0 M4	71.00	2.0 E12	90.00	1.0 M2	67.00	1.0 E9	83.00	2.0 E15	82.00	2.0 E2	92.00	2.0 E12	90.00	1.0 E6	89.00	1.0 M2	67.00
2.0 E12	90.00	2.0 E1	83.00	2.0 E13	82.00	2.0 E10	83.00	2.0 E10	83.00	2.0 E22	76.00	2.0 E20	80.00	2.0 E14	90.00	1.0 E7	78.00	2.0 E12	90.00
2.0 E15	82.00	2.0 E11	82.50	2.0 E17	79.00	2.0 E12	90.00	2.0 E11	82.50	2.0 E5	74.00	2.0 E11	78.00	2.0 E15	82.00	1.0 M1	63.00	2.0 E13	79.00
2.0 E16	83.00	2.0 E12	90.00	2.0 E19	86.00	2.0 E2	92.00	2.0 E14	90.00	2.0 E6	80.00	2.0 E8	81.00	2.0 E16	83.00	1.0 M2	67.00	2.0 E2	92.00
2.0 E22	76.00	2.0 E14	90.00	1.0 E6	80.00	2.0 E6	80.00	2.0 E16	83.00	2.0 E9	89.00	2.0 H1	46.25	2.0 E19	86.00	2.0 E18	81.00	2.0 E6	80.00
2.0 E6	80.00	2.0 E17	79.00	2.0 E7	75.00	2.0 M3	67.00	2.0 E20	80.00	2.0 H1	46.25	2.0 M5	68.00	2.0 E2	92.00	2.0 E19	86.00	2.0 E8	81.00
2.0 H1	46.25	2.0 E3	76.00	2.0 H1	46.25	2.0 M4	53.00	2.0 E8	81.00	2.0 M1	63.00	2.0 M7	66.25	2.0 E5	74.00	2.0 E22	76.00	2.0 H1	46.25
2.0 M10	56.25	2.0 E5	74.00	2.0 M4	53.00	2.0 M6	53.75	2.0 H1	46.25	2.0 M6	53.75	3.0 E12	85.00	2.0 E7	75.00	2.0 E3	76.00	2.0 M10	56.25
2.0 M2	48.75	2.0 E7	75.00	3.0 E10	85.00	2.0 M7	66.25	2.0 M5	68.00	2.0 M7	66.25	3.0 E3	84.00	2.0 E9	89.00	2.0 E5	74.00	2.0 M2	48.75
2.0 M7	66.25	2.0 M7	66.25	3.0 E13	73.00	3.0 E12	85.00	2.0 M9	70.00	2.0 M9	70.00	3.0 E4	74.00	2.0 M2	48.75	2.0 M4	53.00	2.0 M7	66.25
3.0 E12	85.00	3.0 E1	90.00	3.0 E7	83.00	3.0 E15	73.00	3.0 E13	72.00	3.0 E14	83.00	3.0 E5	79.00	2.0 M5	68.00	2.0 M8	52.50	3.0 E1	90.00
3.0 E14	83.00	3.0 E12	85.00	3.0 E8	72.00	3.0 E16	75.00	3.0 E13	90.00	3.0 E3	79.00	3.0 E6	73.00	2.0 M6	53.75	3.0 E10	85.00	3.0 E4	74.00
3.0 E14	90.00	3.0 E14	83.00	3.0 M1	57.50	3.0 E9	84.00	3.0 E2	87.00	3.0 E6	73.00	3.0 E8	72.00	3.0 E10	85.00	3.0 E12	85.00	3.0 E8	72.00
3.0 E6	73.00	3.0 E5	79.00	3.0 M2	61.00	3.0 E6	73.00	3.0 E3	84.00	3.0 E8	72.00	3.0 E9	89.00	3.0 E12	85.00	3.0 E15	73.00	3.0 M1	57.50
3.0 E9	89.00	3.0 E7	83.00	3.0 M3	57.50	3.0 E7	83.00	3.0 M2	61.00	3.0 E9	89.00	3.0 M3	57.50	3.0 E14	83.00	3.0 E5	79.00	3.0 M2	61.00
Difficulty	75.23	Difficulty	79.44	Difficulty	74.54	Difficulty	75.35	Difficulty	78.79	Difficulty	73.31	Difficulty	76.60	Difficulty	79.70	Difficulty	76.03	Difficulty	73.00
Easy	14	Easy	16	Easy	15	Easy	14	Easy	16	Easy	13	Easy	16	Easy	17	Easy	16	Easy	12
Moderate	5	Moderate	4	Moderate	4	Moderate	6	Moderate	3	Moderate	6	Moderate	3	Moderate	3	Moderate	4	Moderate	7
Hard	1	Hard	0	Hard	1	Hard	0	Hard	1	Hard	1	Hard	1	Hard	0	Hard	0	Hard	1
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	5	Topic 1	6	Topic 1	4	Topic 1	6	Topic 1	6	Topic 1	5	Topic 1	2	Topic 1	3	Topic 1	9	Topic 1	6
Topic 2	10	Topic 2	9	Topic 2	9	Topic 2	8	Topic 2	9	Topic 2	10	Topic 2	10	Topic 2	14	Topic 2	7	Topic 2	9
Topic 3	5	Topic 3	5	Topic 3	7	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	8	Topic 3	3	Topic 3	4	Topic 3	5

Table 21 - Experiment #2A - Item Selection - Attempts 11 - 20

Attempt 21		Attempt 22		Attempt 23		Attempt 24		Attempt 25		Attempt 26		Attempt 27		Attempt 28		Attempt 29		Attempt 30	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E14	79.00	1.0 E12	75.00	1.0 E13	76.00	1.0 E13	76.00	1.0 E12	75.00	1.0 E1	80.00	1.0 E11	78.00	1.0 E13	78.00	1.0 E10	77.00	1.0 E13	76.00
1.0 E3	77.00	1.0 E13	76.00	1.0 E1	80.00	1.0 E2	75.00	1.0 E4	76.00	1.0 E9	83.00	1.0 E4	76.00	1.0 E4	76.00	1.0 E11	78.00	1.0 E2	75.00
1.0 E4	76.00	1.0 E4	76.00	1.0 E3	77.00	1.0 E7	78.00	2.0 E16	83.00	2.0 E12	90.00	1.0 E5	94.00	1.0 E5	94.00	1.0 E2	75.00	1.0 E4	76.00
1.0 E5	94.00	1.0 E7	78.00	1.0 E5	94.00	1.0 M1	83.00	2.0 E17	79.00	2.0 E15	82.00	1.0 E8	91.00	1.0 E6	89.00	1.0 E4	76.00	1.0 E5	94.00
1.0 E6	89.00	1.0 E9	83.00	1.0 M4	71.00	2.0 E10	83.00	2.0 E18	81.00	2.0 E4	75.00	2.0 E21	78.00	1.0 M1	83.00	1.0 E8	91.00	1.0 E9	83.00
1.0 E7	78.00	2.0 E13	79.00	2.0 E10	83.00	2.0 E18	81.00	2.0 E4	75.00	2.0 E21	78.00	1.0 M1	83.00	2.0 E13	79.00	1.0 M1	83.00	2.0 E13	79.00
1.0 E8	72.00	2.0 E17	79.00	2.0 E20	80.00	2.0 E2	92.00	2.0 M1	63.00	2.0 E3	76.00	2.0 E13	79.00	2.0 E14	90.00	1.0 M2	67.00	2.0 E17	79.00
1.0 E9	89.00	2.0 E22	76.00	2.0 E21	78.00	2.0 E20	80.00	2.0 M10	56.25	2.0 E6	80.00	2.0 E17	79.00	2.0 E17	79.00	1.0 M4	71.00	2.0 E2	92.00
2.0 E13	79.00	2.0 E3	76.00	2.0 E5	74.00	2.0 E22	75.00	2.0 M2	48.75	2.0 E7	75.00	2.0 E18	81.00	2.0 E5	74.00	2.0 E12	90.00	2.0 E21	78.00
2.0 E15	82.00	2.0 E5	74.00	2.0 E7	75.00	2.0 E4	75.00	2.0 M3	67.00	2.0 E9	89.00	2.0 E19	86.00	2.0 E8	81.00	2.0 E19	86.00	2.0 E9	89.00
2.0 E16	89.00	2.0 E8	81.00	2.0 M10	56.25	2.0 E5	74.00	2.0 M5	68.00	2.0 H1	46.25	2.0 E2	92.00	2.0 H1	46.25	2.0 E21	78.00	2.0 H1	46.25
2.0 E22	76.00	2.0 M1	63.00	2.0 M3	48.75	2.0 E8	81.00	2.0 M7	66.25	2.0 M5	68.00	2.0 E9	76.00	2.0 M6	53.75	2.0 E3	76.00	2.0 M3	48.75
2.0 E5	74.00	2.0 E11	84.00	2.0 M8	52.50	2.0 H1	46.25	2.0 M9	70.00	2.0 M7	66.25	2.0 H1	46.25	2.0 M9	70.00	2.0 E4	75.00	2.0 E1	90.00
2.0 E7	75.00	2.0 E17	72.00	2.0 E11	84.00	2.0 M2	48.75	2.0 E1	90.00	2.0 E10	85.00	2.0 M10	56.25	2.0 E1	90.00	2.0 E9	89.00	2.0 E15	72.00
3.0 E11	84.00	3.0 E4	74.00	3.0 E13	72.00	2.0 M7	66.25	3.0 E11	84.00	3.0 E13	72.00	2.0 M2	48.75	3.0 E13	72.00	2.0 M8	52.50	3.0 E13	75.00
3.0 E14	83.00	3.0 E5	79.00	3.0 E3	84.00	2.0 M9	70.00	3.0 E12	85.00	3.0 E14	83.00	2.0 M8	52.50	3.0 E5	79.00	3.0 E12	85.00	3.0 E16	75.00
3.0 E17	72.00	3.0 E6	73.00	3.0 E7	83.00	3.0 E10	85.00	3.0 E13	72.00	3.0 E15	73.00	3.0 E14	83.00	3.0 E6	73.00	3.0 E16	75.00	3.0 E4	74.00
3.0 E2	87.00	3.0 E9	89.00	3.0 E9	89.00	3.0 E12	85.00	3.0 E16	75.00	3.0 E16	75.00	3.0 E15	73.00	3.0 E7	83.00	3.0 E2	87.00	3.0 E6	73.00
3.0 E9	89.00	3.0 M2	61.00	3.0 M1	57.50	3.0 E6	73.00	3.0 E17	72.00	3.0 E2	87.00	3.0 E6	73.00	3.0 E8	72.00	3.0 E8	72.00	3.0 E8	72.00
3.0 M2	61.00	3.0 M3	57.50	3.0 M2	61.00	3.0 E9	89.00	3.0 E5	84.00	3.0 E4	74.00	3.0 M2	61.00	3.0 M1	57.50	3.0 M1	57.50	3.0 M1	57.50
Difficulty	79.95	Difficulty	75.28	Difficulty	73.80	Difficulty	74.86	Difficulty	73.36	Difficulty	77.18	Difficulty	73.58	Difficulty	76.38	Difficulty	75.65	Difficulty	75.34
Easy	19	Easy	17	Easy	14	Easy	15	Easy	15	Easy	17	Easy	14	Easy	16	Easy	15	Easy	16
Moderate	1	Moderate	9	Moderate	6	Moderate	4	Moderate	7	Moderate	2	Moderate	5	Moderate	3	Moderate	3	Moderate	3
Hard	0	Hard	0	Hard	0	Hard	1	Hard	0	Hard	1	Hard	1	Hard	1	Hard	0	Hard	1
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	8	Topic 1	5	Topic 1	5	Topic 1	4	Topic 1	2	Topic 1	2	Topic 1	6	Topic 1	5	Topic 1	8	Topic 1	5
Topic 2	6	Topic 2	7	Topic 2	8	Topic 2	12	Topic 2	11	Topic 2	11	Topic 2	10	Topic 2	8	Topic 2	7	Topic 2	7
Topic 3	6	Topic 3	8	Topic 3	7	Topic 3	4	Topic 3	7	Topic 3	7	Topic 3	4	Topic 3	7	Topic 3	3	Topic 3	8

Table 22 - Experiment #2A - Item Selection - Attempts 21 - 30

The item selection difficulty varied widely with each iteration (see ‘Difficulty’ columns in tables 20, 21, and 22). The target cut score/difficulty was 76.13. The randomization produced a difficulty range between 73.00 and 79.95 with average of 75.87. The standard deviation of the scores was 2.17 with a 95% confidence interval of 0.778 which means that the true population mean is between 75.1 and 76.64 of the 30 samples. The kurtosis of the average difficulty is -0.527 and the skewness is 0.613. The number of items at each difficulty level from each topic varied with each iteration. Table 23 provides a summary of the statistics for the sample. Figure 7 illustrates the standard distribution curve of the sample.

Sample Difficulty Statistics	
Target Cut Score	76.13
Mean difficulty	75.87
Median	75.34
Minimum	73.00
Maximum	79.95
Variance Target vs. Mean	0.03
Standard Deviation all Averages	2.17
95% Confidence Score	0.777910235
Kurtosis	-0.52653425
Skewness	0.613319589

Table 23 - Difficulty Statistics for Experiment #2A

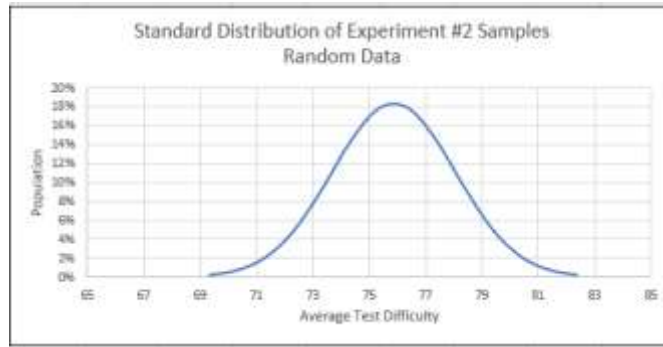


Figure 7 - Standard Distribution of Test Difficulty - Experiment #2A - Random Selection

The content (topic) coverage was erratic as illustrated by the three-color (delineated by sections) display and 'Total From Topic' columns in tables 20, 21, and 22.

Conclusion: All topics were not covered equally in either difficulty or content.

Experiment #2B – Stratified Randomization – Real Client #1 Data

Using the same data from experiment #2A, following the recommended test design, the assessment design function of Questionmark OnDemand was instructed to select 20 items in a stratified random fashion from the sub-topics, following the recommended design shown in table 19 (figure 8).

Question selections	
4 random question(s) from topic 'FAIRNESS RESEARCH 2/1.0 TOPIC 1/1.0 EASY' excluding subtopics (Avoid previously delivered)	
1 random question(s) from topic 'FAIRNESS RESEARCH 2/1.0 TOPIC 1/1.0 MODERATE' excluding subtopics (Avoid previously delivered)	
6 random question(s) from topic 'FAIRNESS RESEARCH 2/2.0 TOPIC 2/2.0 EASY' excluding subtopics (Avoid previously delivered)	
3 random question(s) from topic 'FAIRNESS RESEARCH 2/2.0 TOPIC 2/2.0 MODERATE' excluding subtopics (Avoid previously delivered)	
1 random question(s) from topic 'FAIRNESS RESEARCH 2/2.0 TOPIC 2/2.0 HARD' excluding subtopics (Avoid previously delivered)	
4 random question(s) from topic 'FAIRNESS RESEARCH 2/3.0 TOPIC 3/3.0 EASY' excluding subtopics (Avoid previously delivered)	
1 random question(s) from topic 'FAIRNESS RESEARCH 2/3.0 TOPIC 3/3.0 MODERATE' excluding subtopics (Avoid previously delivered)	

Figure 8 - Item Selection Criteria - Experiment #2B

Thirty (n=30) iterations of a 20-question test were generated as illustrated in tables 24, 25, and 26

Experiment #2 - Directed Random Selection of 20 Items from all 3 topics. Real Client Data. Desired target difficulty is 76.13.																			
Attempt 1		Attempt 2		Attempt 3		Attempt 4		Attempt 5		Attempt 6		Attempt 7		Attempt 8		Attempt 9		Attempt 10	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E13	76.00	1.0 E13	76.00	1.0 E2	75.00	1.0 E10	77.00	1.0 E6	91.00	1.0 E1	80.00	1.0 E9	83.00	1.0 E12	75.00	1.0 E7	78.00	1.0 E13	76.00
1.0 E1	80.00	1.0 E9	83.00	1.0 E12	75.00	1.0 E1	80.00	1.0 E5	94.00	1.0 E11	78.00	1.0 E12	75.00	1.0 E1	80.00	1.0 E4	76.00	1.0 E1	80.00
1.0 E8	91.00	1.0 E14	79.00	1.0 E8	91.00	1.0 E11	78.00	1.0 E14	79.00	1.0 E7	78.00	1.0 E7	78.00	1.0 E7	78.00	1.0 E5	94.00	1.0 E8	91.00
1.0 E4	76.00	1.0 E1	80.00	1.0 E3	77.00	1.0 E3	77.00	1.0 E4	76.00	1.0 E9	83.00	1.0 E11	78.00	1.0 E8	91.00	1.0 E13	76.00	1.0 E3	77.00
1.0 M1	63.00	1.0 M1	63.00	1.0 M2	67.00	1.0 M2	67.00	1.0 M2	67.00	1.0 M3	69.00	1.0 M4	71.00	1.0 M3	69.00	1.0 M4	71.00	1.0 M4	71.00
2.0 E15	82.00	2.0 E22	76.00	2.0 E3	76.00	2.0 E9	89.00	2.0 E11	82.50	2.0 E15	82.00	2.0 E30	80.00	2.0 E7	75.00	2.0 E15	82.00	2.0 E9	89.00
2.0 E14	90.00	2.0 E17	79.00	2.0 E7	75.00	2.0 E13	79.00	2.0 E16	85.00	2.0 E7	75.00	2.0 E6	80.00	2.0 E18	81.00	2.0 E13	79.00	2.0 E4	75.00
2.0 E8	81.00	2.0 E13	79.00	2.0 E21	78.00	2.0 E18	81.00	2.0 E10	85.00	2.0 E9	89.00	2.0 E2	92.00	2.0 E14	90.00	2.0 E17	79.00	2.0 E1	83.00
2.0 E2	92.00	2.0 E9	89.00	2.0 E17	79.00	2.0 E12	90.00	2.0 E17	79.00	2.0 E14	90.00	2.0 E9	89.00	2.0 E20	80.00	2.0 E5	74.00	2.0 E5	74.00
2.0 E18	81.00	2.0 E7	75.00	2.0 E10	83.00	2.0 E8	81.00	2.0 E13	79.00	2.0 E19	86.00	2.0 E15	82.00	2.0 E15	82.00	2.0 E7	75.00	2.0 E13	82.00
2.0 E17	79.00	2.0 E15	82.00	2.0 E14	90.00	2.0 E5	74.00	2.0 E22	76.00	2.0 E4	75.00	2.0 E13	79.00	2.0 E4	75.00	2.0 E4	75.00	2.0 E2	92.00
2.0 M5	68.00	2.0 M9	70.00	2.0 M4	53.00	2.0 M7	66.25	2.0 M7	66.25	2.0 M4	53.00	2.0 M9	70.00	2.0 M9	70.00	2.0 M9	70.00	2.0 M10	56.25
2.0 M10	56.25	2.0 M8	52.50	2.0 M2	48.75	2.0 M2	48.75	2.0 M10	56.25	2.0 M9	70.00	2.0 M7	66.25	2.0 M10	56.25	2.0 M10	56.25	2.0 M9	70.00
2.0 M1	63.00	2.0 M6	53.75	2.0 M6	53.75	2.0 M3	67.00	2.0 M9	70.00	2.0 M10	56.25	2.0 M2	48.75	2.0 M6	53.75	2.0 M1	63.00	2.0 M6	53.75
2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25
3.0 E9	89.00	3.0 E1	90.00	3.0 E10	85.00	3.0 E6	73.00	3.0 E10	85.00	3.0 E13	72.00	3.0 E10	85.00	3.0 E11	84.00	3.0 E17	72.00	3.0 E16	75.00
3.0 E17	72.00	3.0 E10	85.00	3.0 E12	85.00	3.0 E4	74.00	3.0 E4	74.00	3.0 E14	83.00	3.0 E7	83.00	3.0 E5	79.00	3.0 E3	84.00	3.0 E7	83.00
3.0 E11	84.00	3.0 E14	83.00	3.0 E11	84.00	3.0 E17	72.00	3.0 E11	84.00	3.0 E9	89.00	3.0 E17	72.00	3.0 E13	72.00	3.0 E12	85.00	3.0 E5	79.00
3.0 E8	72.00	3.0 E2	87.00	3.0 E1	90.00	3.0 E5	79.00	3.0 E8	72.00	3.0 E3	84.00	3.0 E3	84.00	3.0 E6	79.00	3.0 E10	85.00	3.0 E3	84.00
3.0 M3	57.50	3.0 M1	57.50	3.0 M2	61.00	3.0 M1	57.50	3.0 M2	61.00	3.0 M3	57.50	3.0 M1	57.50	3.0 M2	61.00	3.0 M3	57.50	3.0 M1	57.50
Difficulty	74.95	Difficulty	74.30	Difficulty	73.64	Difficulty	72.84	Difficulty	75.21	Difficulty	74.80	Difficulty	74.99	Difficulty	73.56	Difficulty	73.90	Difficulty	74.74
Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14
Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5
Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5
Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10
Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5

Table 24 - Experiment #2B Item Selections - Attempts 1 - 10

Attempt 11		Attempt 12		Attempt 13		Attempt 14		Attempt 15		Attempt 16		Attempt 17		Attempt 18		Attempt 19		Attempt 20	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E7	78.00	1.0 E1	80.00	1.0 E4	76.00	1.0 E7	78.00	1.0 E12	75.00	1.0 E14	79.00	1.0 E5	94.00	1.0 E5	94.00	1.0 E8	91.00	1.0 E9	83.00
1.0 E2	75.00	1.0 E8	91.00	1.0 E9	83.00	1.0 E9	83.00	1.0 E4	76.00	1.0 E1	80.00	1.0 E7	78.00	1.0 E13	76.00	1.0 E7	78.00	1.0 E8	91.00
1.0 E11	78.00	1.0 E7	78.00	1.0 E5	77.00	1.0 E11	78.00	1.0 E9	83.00	1.0 E8	91.00	1.0 E13	76.00	1.0 E4	76.00	1.0 E2	75.00	1.0 E5	94.00
1.0 E10	77.00	1.0 E6	89.00	1.0 E6	89.00	1.0 E3	77.00	1.0 E13	76.00	1.0 E7	78.00	1.0 E4	76.00	1.0 E2	75.00	1.0 E11	78.00	1.0 E2	75.00
1.0 M3	69.00	1.0 M4	71.00	1.0 M2	67.00	1.0 M4	71.00	1.0 M3	69.00	1.0 M4	71.00	1.0 M1	63.00	1.0 M1	63.00	1.0 M3	69.00	1.0 M3	69.00
2.0 E12	90.00	2.0 E13	79.00	2.0 E5	74.00	2.0 E7	75.00	2.0 E8	81.00	2.0 E3	76.00	2.0 E12	90.00	2.0 E22	76.00	2.0 E8	81.00	2.0 E10	83.00
2.0 E7	75.00	2.0 E12	90.00	2.0 E16	83.00	2.0 E15	82.00	2.0 E12	90.00	2.0 E14	90.00	2.0 E17	79.00	2.0 E15	82.00	2.0 E17	79.00	2.0 E11	82.50
2.0 E21	78.00	2.0 E21	78.00	2.0 E15	79.00	2.0 E13	79.00	2.0 E11	82.50	2.0 E5	74.00	2.0 E5	74.00	2.0 E7	75.00	2.0 E21	78.00	2.0 E20	80.00
2.0 E10	83.00	2.0 E15	82.00	2.0 E1	83.00	2.0 E9	89.00	2.0 E19	86.00	2.0 E6	80.00	2.0 E20	80.00	2.0 E19	86.00	2.0 E15	86.00	2.0 E15	82.00
2.0 E22	76.00	2.0 E2	92.00	2.0 E17	79.00	2.0 E5	74.00	2.0 E20	80.00	2.0 E14	83.00	2.0 E10	83.00	2.0 E9	89.00	2.0 E4	75.00	2.0 E1	83.00
2.0 E15	82.00	2.0 E19	86.00	2.0 E14	90.00	2.0 E16	83.00	2.0 E18	81.00	2.0 E22	76.00	2.0 E19	86.00	2.0 E8	81.00	2.0 E10	83.00	2.0 E3	76.00
2.0 M5	88.00	2.0 M10	56.25	2.0 M5	88.00	2.0 M10	56.25	2.0 M5	88.00	2.0 M9	70.00	2.0 M10	56.25	2.0 M7	66.25	2.0 M4	53.00	2.0 M5	88.00
2.0 M9	70.00	2.0 M8	52.50	2.0 M3	67.00	2.0 M4	53.00	2.0 M10	56.25	2.0 M4	53.00	2.0 M4	53.00	2.0 M6	53.75	2.0 M1	63.00	2.0 M2	48.75
2.0 M2	48.75	2.0 M6	53.75	2.0 M9	70.00	2.0 M9	70.00	2.0 M2	48.75	2.0 M2	48.75	2.0 M9	70.00	2.0 M8	52.50	2.0 M10	56.25	2.0 M10	56.25
2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25
3.0 E15	73.00	3.0 E2	87.00	3.0 E17	72.00	3.0 E4	74.00	3.0 E7	83.00	3.0 E3	84.00	3.0 E17	72.00	3.0 E16	75.00	3.0 E6	79.00	3.0 E13	72.00
3.0 E14	83.00	3.0 E14	83.00	3.0 E13	72.00	3.0 E13	72.00	3.0 E17	72.00	3.0 E14	83.00	3.0 E5	79.00	3.0 E16	90.00	3.0 E13	85.00	3.0 E4	74.00
3.0 E3	84.00	3.0 E13	72.00	3.0 E14	83.00	3.0 E9	89.00	3.0 E16	75.00	3.0 E2	87.00	3.0 E6	73.00	3.0 E13	72.00	3.0 E11	84.00	3.0 E11	84.00
3.0 E9	89.00	3.0 E10	85.00	3.0 E8	72.00	3.0 E15	73.00	3.0 E9	89.00	3.0 E17	72.00	3.0 E2	87.00	3.0 E12	85.00	3.0 E2	87.00	3.0 E10	85.00
3.0 M3	57.50	3.0 M3	57.50	3.0 M3	57.50	3.0 M3	57.50	3.0 M3	57.50	3.0 M3	57.50	3.0 M3	57.50	3.0 M2	61.00	3.0 M2	61.00	3.0 M3	57.50
Difficulty	74.03	Difficulty	75.46	Difficulty	74.39	Difficulty	73.00	Difficulty	74.83	Difficulty	75.98	Difficulty	73.85	Difficulty	73.24	Difficulty	74.74	Difficulty	73.76
Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14
Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5
Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5
Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10
Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5

Table 25 - Experiment #2B Item Selections - Attempts 11 - 20

Attempt 21		Attempt 22		Attempt 23		Attempt 24		Attempt 25		Attempt 26		Attempt 27		Attempt 28		Attempt 29		Attempt 30	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E13	76.00	1.0 E10	77.00	1.0 E5	94.00	1.0 E4	76.00	1.0 E6	89.00	1.0 E5	94.00	1.0 E9	83.00	1.0 E11	78.00	1.0 E14	79.00	1.0 E4	76.00
1.0 E3	77.00	1.0 E3	77.00	1.0 E13	76.00	1.0 E7	78.00	1.0 E11	78.00	1.0 E12	75.00	1.0 E5	94.00	1.0 E8	91.00	1.0 E10	77.00	1.0 E8	91.00
1.0 E2	75.00	1.0 E13	76.00	1.0 E3	77.00	1.0 E5	94.00	1.0 E9	83.00	1.0 E3	77.00	1.0 E1	80.00	1.0 E9	83.00	1.0 E11	78.00	1.0 E14	79.00
1.0 E4	76.00	1.0 E9	83.00	1.0 E2	75.00	1.0 E11	78.00	1.0 E13	76.00	1.0 E8	91.00	1.0 E8	91.00	1.0 E4	76.00	1.0 E2	75.00	1.0 E9	83.00
1.0 M4	71.00	1.0 M3	69.00	1.0 M1	61.00	1.0 M2	67.00	1.0 M4	71.00	1.0 M1	63.00	1.0 M3	69.00	1.0 M1	63.00	1.0 M2	67.00	1.0 M2	67.00
2.0 E10	83.00	2.0 E4	75.00	2.0 E4	75.00	2.0 E22	76.00	2.0 E10	83.00	2.0 E3	76.00	2.0 E21	78.00	2.0 E5	76.00	2.0 E22	76.00	2.0 E19	86.00
2.0 E4	75.00	2.0 E2	92.00	2.0 E13	79.00	2.0 E19	86.00	2.0 E8	81.00	2.0 E16	83.00	2.0 E22	76.00	2.0 E22	76.00	2.0 E14	90.00	2.0 E4	75.00
2.0 E2	92.00	2.0 E5	74.00	2.0 E18	81.00	2.0 E13	79.00	2.0 E20	80.00	2.0 E4	75.00	2.0 E5	74.00	2.0 E17	79.00	2.0 E19	86.00	2.0 E1	83.00
2.0 E10	83.00	2.0 E7	75.00	2.0 E15	82.00	2.0 E10	83.00	2.0 E6	80.00	2.0 E14	90.00	2.0 E4	75.00	2.0 E16	83.00	2.0 E8	81.00	2.0 E2	92.00
2.0 E5	74.00	2.0 E12	90.00	2.0 E6	80.00	2.0 E7	75.00	2.0 E12	90.00	2.0 E10	83.00	2.0 E13	79.00	2.0 E7	75.00	2.0 E3	76.00	2.0 E15	82.00
2.0 E10	80.00	2.0 E10	83.00	2.0 E8	81.00	2.0 E12	90.00	2.0 E1	83.00	2.0 E1	83.00	2.0 E15	82.00	2.0 E9	89.00	2.0 E5	74.00	2.0 E14	90.00
2.0 M1	63.00	2.0 M1	63.00	2.0 M10	56.25	2.0 M10	56.25	2.0 M10	56.25	2.0 M10	56.25	2.0 M7	66.25	2.0 M7	66.25	2.0 M2	67.00	2.0 M3	67.00
2.0 M5	68.00	2.0 M5	68.00	2.0 M9	70.00	2.0 M7	66.25	2.0 M4	53.00	2.0 M6	53.75	2.0 M1	63.00	2.0 M2	48.75	2.0 M2	48.75	2.0 M5	68.00
2.0 M6	53.75	2.0 M4	53.00	2.0 M3	67.00	2.0 M6	53.75	2.0 M6	53.75	2.0 M4	53.00	2.0 M1	63.00	2.0 M2	48.75	2.0 M5	68.00	2.0 M8	52.50
2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25	2.0 H1	46.25
3.0 E14	83.00	3.0 E3	84.00	3.0 E9	89.00	3.0 E2	87.00	3.0 E13	72.00	3.0 E10	85.00	3.0 E17	72.00	3.0 E13	72.00	3.0 E11	84.00	3.0 E16	75.00
3.0 E15	73.00	3.0 E1	90.00	3.0 E3	84.00	3.0 E16	75.00	3.0 E7	83.00	3.0 E13	72.00	3.0 E8	73.00	3.0 E10	85.00	3.0 E15	72.00	3.0 E6	73.00
3.0 E7	83.00	3.0 E6	73.00	3.0 E5	79.00	3.0 E10	85.00	3.0 E17	72.00	3.0 E1	90.00	3.0 E16	75.00	3.0 E7	83.00	3.0 E1	90.00	3.0 E9	89.00
3.0 E6	73.00	3.0 E11	84.00	3.0 E13	72.00	3.0 E7	83.00	3.0 E5	79.00	3.0 E8	73.00	3.0 E2	87.00	3.0 E14	83.00	3.0 E2	87.00	3.0 E14	83.00
3.0 M3	57.50	3.0 M2	61.00	3.0 M3	57.50	3.0 M1	57.50	3.0 M3	57.50	3.0 M1	57.50	3.0 M3	57.50	3.0 M2	61.00	3.0 M3	57.50	3.0 M3	57.50
Difficulty	73.13	Difficulty	74.66	Difficulty	74.20	Difficulty	74.60	Difficulty	73.34	Difficulty	73.84	Difficulty	73.74	Difficulty	73.86	Difficulty	73.98	Difficulty	75.76
Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14	Easy	14
Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5	Moderate	5
Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1	Hard	1
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5	Topic 1	5
Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10	Topic 2	10
Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5	Topic 3	5

Table 26 - Experiment #2B Item Selections - Attempts 21 - 30

The item selection difficulty remained constant as stratified with each iteration (see ‘Difficulty’ columns in tables 24, 25, and 26). The target cut score/difficulty was 76.13. The stratified randomization produced a difficulty range between 73.00 and 75.76 with average (mean) of 74.11. The standard deviation of the scores was 0.74 with a 95% confidence interval of 0.2635 which means that the true population mean is between 73.85 and 74.37 of the 30 samples. The kurtosis of the average difficulty is 0.117 and the skewness is 0.579. The number of items at each difficulty level

from each topic varied with each iteration. Table 27 provides a summary of the statistics for the sample. Figure 9 illustrates the standard distribution curve of the sample.

Sample Difficulty Statistics	
Target Cut Score	76.13
Mean difficulty	74.11
Median	73.98
Minimum	73.00
Maximum	75.76
Variance Target vs. Mean	2.04
Standard Deviation all Averages	0.74
95% Confidence Score	0.263545877
Kurtosis	0.117166773
Skewness	0.579229905

Table 27 - Difficulty Statistics for Experiment #2B

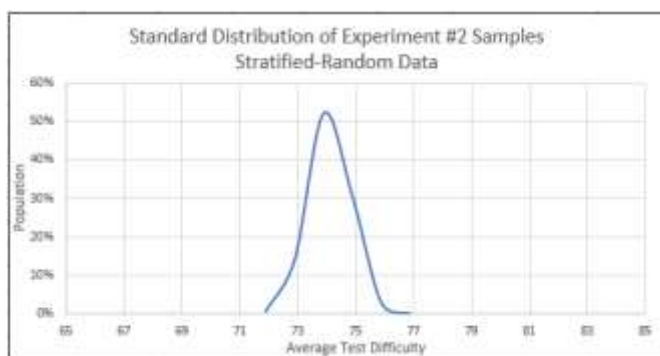


Figure 9 - Standard Distribution of Test Difficulty - Experiment #2B - Stratified Random Selection

The content (topic) coverage was equal as stratified for each iteration as illustrated by the three-color (delineated by sections) display and 'Total From Topic' columns in tables 18A, 18B, and 18C.

Conclusion: All topics were covered equally as desired in both difficulty and content. Comparing the distribution of test difficulty scores in figures 7 and 9 shows that the stratified randomization consistently produced tests well within an acceptable range to meet the desired cut score of 76.13. Although the simple randomization in experiment 2A produced tests with an average (mean) difficulty (75.87) closer to the desired difficulty (76.13) than the stratified randomization in experiment 2B (74.11), the standard deviation of the results in experiment 2B (0.74) indicated significantly less variance from attempt to attempt than the standard deviation produced by experiment 2A (2.17).

Experiment #3A – Random Selection – Real Client #2 Data

Using cut score (Parry Method) results from real client data (Tables 28, 29, and 30 – partial views), the items were divided into three topics within the main topic and further subdivided into difficulty sub-sub topics based upon the results of the cut score rating tool. The cut score/difficulty for the entire database (134 items) was determined to be 64.37% by averaging all three topic cut scores.

CUT SCORE CALCULATION TOOL

Course/Certification Name:

FAIRNESS RESEARCH 3

Test Name:

TEST NAME

Facilitator Name/Date:

Facilitator Name/Date

Revision 1 Facilitator Name/Date

Revision 2 Facilitator Name/Date

Date: mm/dd/yyyy

Enter Topic/TPQ/Subject ID:

Topic 1

This spreadsheet tool is the intellectual property of and Copyright ©2018-2019 by Compass Consultants, LLC. Use is limited to the terms of the End User License Agreement (EULA). This copy is licensed for: 30 DAY DEMO ONLY

Test Item QID	Enter * If New, R If Retired	Difficulty Metatag	Average Percentage Correct (Angoff Rating)	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name	Expert 6 Name	Expert 7 Name	Expert 8 Name	Expert 9 Name	Expert 10 Name	Standard Deviation
1.0 M1		Moderate	48.57	60	60	40	45	40	55	40				9.45
1.0 M2		Moderate	62.14	55	70	60	75	65	60	50				8.59
1.0 M3		Moderate	57.14	50	60	70	60	60	60	40				8.51
	R			50	65	60	65	60	60	40				
1.0 M4		Moderate	62.86	70	70	70	60	60	60	50				7.56
1.0 M5		Moderate	60.00	60	70	70	50	50	70	50				10.00
1.0 E1			77.14	70	85	80	75	70	90	70				8.09
1.0 E2			75.00	70	80	90	75	70	80	60				8.57
1.0 E3			77.14	70	80	90	75	75	80	70				8.99
1.0 E4			84.29	75	90	95	90	75	90	75				8.86
1.0 M6		Moderate	62.14	55	70	50	50	70	70	70				9.94
1.0 H1		Hard	41.43	30	50	35	35	50	50	40				8.52
1.0 M7		Moderate	65.57	60	65	70	50	65	65	70				8.90
1.0 M8		Moderate	55.00	55	60	65	50	50	65	40				8.13
	R			70	80	75	80	70	90	40				
1.0 M9		Moderate	57.14	50	50	70	50	50	70	60				8.51
1.0 M10		Moderate	48.57	50	45	50	35	55	55	50				8.90
1.0 M11		Moderate	57.86	50	70	50	50	65	60	60				8.09

Topic Cut Score

58.00

Moderate Difficulty

Approximate Difficulty Rating

25 - 48.5 : Hard

48.4 - 71.7 : Moderate

71.8 - 95 : Easy

Standard Deviation

A standard deviation of more than 10 will trigger an alert. Discuss the outliers with the judges who set them to determine why. Change as necessary.

7	Easy	In this section	17%
28	Moderate	In this section	67%
7	Hard	In this section	17%
42	TOTAL		100%

Table 28 - Experiment #3A - Difficulty Calculations - Real Data - Topic 1

CUT SCORE CALCULATION TOOL

Course/Certification Name:		FAIRNESS RESEARCH 3				Test Name:		TEST NAME					
Facilitator Name/Date:		Facilitator Name/Date				Revision 1 Facilitator Name/Date				Date: mm/dd/yyyy			
Enter Topic/TPQ/Subject ID:		Topic 2				Revision 2 Facilitator Name/Date							

This spreadsheet tool is the intellectual property of and Copyright ©2018-2019 by Compass Consultants, LLC. Use is limited to the terms of the End User License Agreement (EULA). This copy is licensed for: 30 DAY DEMO ONLY

Test Item QID	Enter * If New, R If Retired	Difficulty Metatag	Average Percentage Correct (Angoff Rating)	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name	Expert 6 Name	Expert 7 Name	Expert 8 Name	Expert 9 Name	Expert 10 Name	Standard Deviation
2.0 E1			75.00	70	80	75	75	80	85	60				8.16
2.0 E2			72.86	65	80	75	65	80	85	60				9.51
2.0 M1		Moderate	60.00	50	65	70	50	50	65	70				9.57
2.0 E3			73.57	60	70	80	75	70	90	70				8.45
2.0 E4			77.14	70	80	85	75	80	90	60				9.94
2.0 E5			74.29	60	85	65	75	80	75	80				8.86
2.0 M2		Moderate	69.29	60	80	75	60	70	80	60				9.32
2.0 E6			73.57	60	70	80	65	75	85	80				9.00
2.0 E7			80.00	70	80	85	85	80	90	70				7.64
2.0 E8			81.43	70	85	95	70	80	90	80				8.45
2.0 M3		Moderate	71.43	65	75	85	65	70	80	60				9.00
2.0 E9			77.14	70	80	85	85	80	90	70				9.06
2.0 M4		Moderate	67.14	65	75	80	50	75	75	50			11.54	
2.0 E10			84.29	75	85	90	95	80	85	80				6.73
2.0 E11			82.14	70	85	90	95	80	85	70				9.51
2.0 E12			74.29	65	70	85	85	75	70	70				7.87
2.0 E13			77.86	70	80	85	80	75	85	70				8.36
2.0 E14			75.00	65	65	80	80	75	80	80				7.07

Topic Cut Score

69.00

Moderate Difficulty

Approximate Difficulty Rating

25 - 48.5 : Hard

48.4 - 71.7 : Moderate

71.8 - 95 : Easy

Standard Deviation

A standard deviation of more than 10 will trigger an alert. Discuss the outliers with the judges who set them to determine why. Change as necessary.

24	Easy	In this section	46%
27	Moderate	In this section	52%
1	Hard	In this section	2%
52	TOTAL		100%

Table 29 - Experiment #3A - Difficulty Calculations - Real Data - Topic 2

CUT SCORE CALCULATION TOOL

Course/Certification Name:

FAIRNESS RESEARCH 3

Test Name:

TEST NAME

Facilitator Name/Date:

Facilitator Name/Date

Revision 1 Facilitator Name/Date

Revision 2 Facilitator Name/Date

Date:

mm/dd/yyyy

Enter Topic/TPO/Subject ID:

Topic 3

This spreadsheet tool is the intellectual property of and Copyright © 2019-2023 by Compass Consultants, LLC. Use is limited to the terms of the Real User License Agreement (RULA). This copy is licensed for 30 DAY DEMO ONLY.

Test Item QID	Enter * If New, R If Retired	Difficulty Metatag	Average Percentage Correct (Angoff Rating)	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name	Expert 6 Name	Expert 7 Name	Expert 8 Name	Expert 9 Name	Expert 10 Name	Standard Deviation
3.0 E1			74.29	60	70	90	75	75	80	70				9.33
3.0 E2			77.14	70	70	90	75	75	90	70				9.06
3.0 E3			77.14	70	70	90	75	75	90	70				9.06
3.0 E4			78.57	70	80	90	75	75	90	70				8.52
3.0 M1		Moderate	67.14	60	65	70	65	70	80	60				6.99
3.0 M2		Moderate	68.57	65	70	75	65	65	65	40				11.07
3.0 M3		Moderate	61.43	65	65	75	50	50	65	60				9.00
3.0 E5			78.57	70	80	90	75	75	90	70				8.52
3.0 E6			79.29	70	80	90	75	70	90	80				8.88
3.0 E7			73.57	65	70	85	70	70	85	70				8.02
3.0 E8			73.57	70	75	90	70	75	75	60				9.30
3.0 E9			77.14	75	75	90	65	75	90	70				9.52
3.0 E10			78.57	75	80	90	75	70	90	70				8.52
3.0 E11			77.86	65	80	90	70	80	90	70				9.94
3.0 M4		Moderate	62.86	65	65	70	50	55	65	70				7.56
3.0 M5		Moderate	71.43	60	70	80	65	65	80	80				8.52
3.0 E12			77.86	70	80	90	75	75	90	85				9.52
3.0 M6		Moderate	58.57	50	50	75	50	50	65	70				11.07

Topic Cut Score

66.00

Moderate Difficulty

Approximate Difficulty Rating

25 - 48.3 Hard

48.4 - 71.7 Moderate

71.8 - 95 Easy

Standard Deviation

A standard deviation of more than 10 will trigger an alert. Discuss the outliers with the judges who set them to determine why. Change as necessary.

14	Easy	In this section	35%
24	Moderate	In this section	60%
2	Hard	In this section	5%
40	TOTAL		100%

Table 30 - Experiment #3A - Difficulty Calculations - Real Data - Topic 3

Topic 1 consisted of 42 test-items with a difficulty rating of 58 (Moderate). Topic 2 consisted of 52 test-items with a difficulty rating of 69 (Moderate). Topic 3 consisted of 40 test-items with a difficulty rating of 66 (Moderate). Each topic had a mix of hard, moderate, and easy items. (Table 31)

Topic	Topic Cut Score & Difficulty	Items in Topic	% of Total Items	Available Hard	% From Topic	Available Mod	% From Topic	Available Easy	% From Topic
Topic 1	58	42	31.34%	7	17%	28	67%	7	17%
Topic 2	69	52	38.81%	1	2%	27	52%	24	46%
Topic 3	66	40	29.85%	2	5%	24	60%	14	35%

Table 31 - Experiment #3 - Item Difficulty Distribution by Topic

Referring to the design philosophy of the Compass Consultants spreadsheet tool as to the number of items drawn from each section, it appears that topic 2 was considered to be the ‘most’ important with 38.81% of the available items, topic 1 was the next ‘most’ important with 31.34% and topic 3 was the least important with 29.85%.

The final test design to maintain fairness in both content and difficulty is shown in table 32.

Total # Needed From Topic	Use Hard (Calculated)	Use Hard (Actual)	Use Mod (Calculated)	Use Mod (Actual)	Use Easy (Calculated)	Use Easy (Actual)	Topic
6.27	1.04	1	4.18	4	1.04	1	Topic 1
7.76	0.15	1	4.03	4	3.58	3	Topic 2
5.97	0.30	1	3.58	3	2.09	2	Topic 3

Table 32 - Experiment #3 - Recommended Test Design

Ignoring the recommendation of item distribution to maintain the cut score, the assessment design function of Questionmark OnDemand was instructed to select 20 items at random from the single topic containing all of the test-items. Tables 33, 34, and 35 present the results.

Experiment 3A - Random Selection of 20 items from all 3 topics. Real Client Data. Desired target difficulty is 64.37																			
Attempt 1		Attempt 2		Attempt 3		Attempt 4		Attempt 5		Attempt 6		Attempt 7		Attempt 8		Attempt 9		Attempt 10	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 E7	72.86	1.0 H4	40.00	1.0 E5	72.86	1.0 E5	72.86	1.0 E1	77.14	1.0 H7	42.14	1.0 E4	84.29	1.0 M11	57.86	1.0 E4	84.29	1.0 E5	72.86
1.0 H1	41.43	1.0 M15	58.57	1.0 M10	48.57	1.0 M1	48.57	1.0 M16	65.00	1.0 M12	65.00	1.0 E5	72.86	1.0 M12	65.00	1.0 M2	62.14	1.0 E7	72.86
1.0 H5	47.86	1.0 M17	60.71	1.0 M11	57.86	1.0 M19	50.00	1.0 M20	59.29	1.0 M13	56.43	1.0 H1	41.43	1.0 M18	50.71	1.0 M22	61.43	1.0 H2	44.29
1.0 H6	32.86	1.0 M2	62.14	1.0 M13	56.43	1.0 M2	62.14	2.0 E14	75.00	1.0 M25	54.29	1.0 H4	40.00	1.0 M23	71.43	2.0 E16	75.71	1.0 H5	47.86
1.0 M12	62.14	1.0 M3	57.14	1.0 M14	53.57	1.0 M5	60.00	2.0 E15	75.71	1.0 M4	62.86	1.0 M12	65.00	2.0 E1	75.00	2.0 E19	75.00	1.0 M2	62.14
1.0 M13	56.43	2.0 E13	77.86	1.0 M28	55.71	1.0 M7	63.57	2.0 E17	77.14	1.0 M6	62.14	1.0 M13	56.43	2.0 E13	77.86	2.0 E21	73.57	1.0 M3	57.14
1.0 M16	65.00	2.0 E23	76.43	1.0 M6	62.14	2.0 E10	84.29	2.0 E24	72.14	2.0 E13	77.86	1.0 M16	65.00	2.0 E18	80.71	2.0 M1	60.00	2.0 E10	84.29
1.0 M18	50.71	2.0 E5	74.29	2.0 E21	73.57	2.0 E21	73.57	2.0 E5	74.29	2.0 E15	75.71	1.0 M18	50.71	2.0 E19	75.00	2.0 M14	67.86	2.0 E14	75.00
2.0 E21	73.57	2.0 M12	60.71	2.0 E9	77.14	2.0 E22	74.29	2.0 E6	73.57	2.0 E17	77.14	1.0 M21	57.86	2.0 M1	60.00	2.0 M24	60.71	2.0 E16	75.71
2.0 E22	74.29	2.0 M13	64.29	2.0 M10	64.29	2.0 E24	72.14	2.0 E7	80.00	2.0 E19	75.00	1.0 M4	62.86	2.0 M10	64.29	2.0 M27	57.86	2.0 E9	77.14
2.0 E7	80.00	2.0 M15	61.43	2.0 M2	69.29	2.0 M10	64.29	2.0 H1	46.43	2.0 M10	64.29	2.0 E21	73.57	2.0 M2	69.29	2.0 M3	71.43	2.0 H1	46.43
2.0 M11	57.14	2.0 M17	67.86	2.0 M21	59.29	2.0 M14	67.86	2.0 M12	60.71	2.0 M16	64.29	2.0 E6	73.57	2.0 M21	59.29	2.0 M5	60.71	2.0 M15	61.43
2.0 M12	60.71	2.0 M23	57.86	2.0 M23	57.86	2.0 M15	61.43	2.0 M16	64.29	2.0 M19	60.71	2.0 E7	80.00	2.0 M22	58.57	2.0 M6	70.00	2.0 M26	57.14
2.0 M5	70.00	2.0 M24	60.71	2.0 M25	57.14	2.0 M18	59.29	3.0 E2	77.14	2.0 M25	60.00	2.0 E8	81.43	2.0 M25	60.00	2.0 M8	67.86	2.0 M27	57.86
3.0 E12	77.86	2.0 M5	60.71	3.0 E7	73.57	2.0 M25	60.00	3.0 E3	77.14	2.0 M3	71.43	2.0 M14	67.86	2.0 M5	60.71	3.0 E10	78.57	3.0 E10	78.57
3.0 E13	72.86	3.0 H2	46.43	3.0 E8	73.57	3.0 H2	46.43	3.0 E4	78.57	3.0 E14	72.14	2.0 M21	59.29	2.0 M7	65.00	3.0 E5	78.57	3.0 M1	67.14
3.0 M19	70.00	3.0 M22	57.86	3.0 M1	67.14	3.0 M12	55.00	3.0 E9	77.14	3.0 E7	73.57	3.0 E3	77.14	2.0 M9	65.00	3.0 M17	60.00	3.0 M18	60.00
3.0 M2	63.57	3.0 M24	54.29	3.0 M12	55.00	3.0 M13	60.00	3.0 M12	55.00	3.0 M15	49.29	3.0 E7	73.57	3.0 E11	77.86	3.0 M21	62.86	3.0 M21	62.86
3.0 M3	61.43	3.0 M5	71.43	3.0 M15	49.29	3.0 M14	60.71	3.0 M3	61.43	3.0 M22	57.86	3.0 M13	60.00	3.0 M13	60.00	3.0 M22	57.86	3.0 M5	71.43
3.0 M5	58.57	3.0 M7	57.14	3.0 M16	68.57	3.0 M17	60.00	3.0 M7	57.14	3.0 M9	56.43	3.0 M8	65.00	3.0 M17	60.00	3.0 M9	56.43	3.0 M6	58.57
Difficulty	62.46	Difficulty	61.59	Difficulty	62.64	Difficulty	62.82	Difficulty	69.21	Difficulty	63.93	Difficulty	65.99	Difficulty	65.88	Difficulty	67.14	Difficulty	64.54
Easy	6	Easy	3	Easy	5	Easy	5	Easy	12	Easy	6	Easy	8	Easy	5	Easy	6	Easy	7
Moderate	11	Moderate	13	Moderate	15	Moderate	14	Moderate	7	Moderate	13	Moderate	10	Moderate	15	Moderate	14	Moderate	10
Hard	3	Hard	2	Hard	0	Hard	1	Hard	1	Hard	1	Hard	2	Hard	0	Hard	0	Hard	3
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	8	Topic 1	5	Topic 1	7	Topic 1	6	Topic 1	3	Topic 1	6	Topic 1	10	Topic 1	4	Topic 1	3	Topic 1	6
Topic 2	6	Topic 2	10	Topic 2	7	Topic 2	9	Topic 2	10	Topic 2	8	Topic 2	6	Topic 2	13	Topic 2	11	Topic 2	8
Topic 3	6	Topic 3	5	Topic 3	6	Topic 3	5	Topic 3	7	Topic 3	5	Topic 3	4	Topic 3	5	Topic 3	6	Topic 3	6

Table 33 - Experiment #3A - Item Selection - Attempts 1 - 10

Attempt 11		Attempt 12		Attempt 13		Attempt 14		Attempt 15		Attempt 16		Attempt 17		Attempt 18		Attempt 19		Attempt 20	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 M13	56.43	1.0 M10	48.57	1.0 M14	53.57	1.0 H7	42.14	1.0 H1	41.43	1.0 E5	72.86	1.0 E4	84.29	1.0 E2	75.00	1.0 E7	72.86	1.0 E4	84.29
1.0 M14	53.57	1.0 M11	57.86	1.0 M17	60.71	1.0 M11	57.86	1.0 H4	40.00	1.0 M14	53.57	1.0 H1	41.43	1.0 H1	41.43	1.0 M15	58.57	1.0 H4	40.00
1.0 M22	61.43	1.0 M16	65.00	1.0 M18	50.71	1.0 M12	65.00	1.0 H5	47.86	1.0 M18	50.71	1.0 H6	32.86	1.0 M10	48.57	1.0 M27	53.57	1.0 M2	62.14
1.0 M27	53.57	1.0 M18	50.71	1.0 M20	59.29	1.0 M19	50.00	1.0 M15	58.57	1.0 M18	48.57	1.0 M12	65.00	1.0 M15	58.57	1.0 M7	63.57	1.0 M22	61.43
1.0 M9	57.14	1.0 M20	59.29	1.0 M21	57.86	1.0 M26	52.86	1.0 M2	62.14	1.0 M27	53.57	1.0 M17	60.71	1.0 M19	50.00	1.0 M9	57.14	1.0 M25	54.29
2.0 E10	84.29	1.0 M11	57.86	1.0 M22	61.43	1.0 M27	53.57	1.0 M3	57.14	2.0 E16	75.71	1.0 M25	54.29	1.0 M2	62.14	2.0 E11	82.14	1.0 M5	60.00
2.0 E2	72.86	2.0 E13	77.86	1.0 M5	60.00	1.0 M27	62.14	2.0 E13	77.86	2.0 E17	77.14	1.0 M6	62.14	1.0 M6	62.14	2.0 E14	75.00	1.0 M8	55.00
2.0 M1	60.00	2.0 E15	75.71	2.0 E14	75.00	1.0 M6	62.14	2.0 E2	72.86	2.0 E23	76.43	2.0 E1	75.00	1.0 M9	57.14	2.0 E18	80.71	1.0 M9	57.14
2.0 M13	64.29	2.0 E23	76.43	2.0 E6	73.57	2.0 E10	84.29	2.0 E22	74.29	2.0 E4	77.14	2.0 E23	76.43	2.0 E22	74.29	2.0 E19	75.00	2.0 E13	77.86
2.0 M15	61.43	2.0 E4	77.14	2.0 E9	77.14	2.0 E12	74.29	2.0 M18	59.29	2.0 E8	73.57	2.0 E9	77.14	2.0 E5	74.29	2.0 M11	57.14	2.0 E6	73.57
2.0 M27	57.86	2.0 E9	77.14	2.0 M14	67.86	2.0 E4	77.14	2.0 M18	60.00	2.0 M12	60.71	2.0 M12	60.71	2.0 M12	60.71	2.0 M11	67.86	2.0 E8	81.43
2.0 M6	70.00	2.0 M23	57.86	2.0 M26	57.14	2.0 M8	67.86	2.0 M2	69.29	2.0 M15	61.43	2.0 M15	61.43	2.0 M21	59.29	2.0 M23	57.86	2.0 M13	64.29
3.0 E12	77.86	2.0 M4	67.14	2.0 M5	60.71	3.0 E12	77.86	2.0 M20	67.14	2.0 M20	67.14	2.0 M26	57.14	2.0 M22	58.57	2.0 M26	57.14	2.0 M14	67.86
3.0 E3	77.14	3.0 E11	77.86	2.0 M7	65.00	3.0 E8	77.14	2.0 M29	57.86	2.0 M24	60.71	2.0 M3	71.43	2.0 M24	60.71	2.0 M5	60.71	2.0 M26	57.14
3.0 E8	73.57	3.0 E11	74.29	3.0 E10	78.57	3.0 H1	45.00	2.0 M27	57.86	2.0 M26	57.14	2.0 M7	65.00	2.0 M5	60.71	2.0 M7	65.00	2.0 M8	67.86
3.0 H2	46.43	3.0 E4	78.57	3.0 E6	79.29	3.0 M12	55.00	2.0 M3	71.43	3.0 E2	77.14	3.0 E12	77.86	2.0 M7	65.00	3.0 E1	74.29	3.0 M14	60.71
3.0 M1	67.14	3.0 E6	79.29	3.0 E7	73.57	3.0 M15	49.29	2.0 M5	60.71	3.0 E7	73.57	3.0 E3	78.57	3.0 E7	73.57	3.0 E4	78.57	3.0 M14	67.14
3.0 M16	68.57	3.0 M14	60.71	3.0 E9	77.14	3.0 M18	60.00	3.0 M13	60.00	3.0 H1	45.00	3.0 M14	60.71	3.0 E8	73.57	3.0 E6	79.29	3.0 M21	62.86
3.0 M20	60.00	3.0 M17	60.00	3.0 M12	55.00	3.0 M20	60.00	3.0 M24	54.29	3.0 M17	60.00	3.0 M16	68.57	3.0 M20	60.00	3.0 H2	46.43	3.0 M22	57.86
3.0 M24	54.29	3.0 M5	71.43	3.0 M13	60.00	3.0 M9	56.43	3.0 M6	58.57	3.0 M2	63.57	3.0 M18	60.00	3.0 M24	54.29	3.0 M4	62.86	3.0 M3	61.43
Difficulty	63.89	Difficulty	67.54	Difficulty	65.18	Difficulty	61.50	Difficulty	60.43	Difficulty	64.28	Difficulty	64.54	Difficulty	61.50	Difficulty	66.29	Difficulty	63.72
Easy	5	Easy	9	Easy	7	Easy	5	Easy	3	Easy	8	Easy	6	Easy	5	Easy	8	Easy	4
Moderate	14	Moderate	11	Moderate	15	Moderate	13	Moderate	14	Moderate	11	Moderate	12	Moderate	14	Moderate	11	Moderate	15
Hard	1	Hard	0	Hard	0	Hard	2	Hard	3	Hard	1	Hard	2	Hard	1	Hard	1	Hard	1
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	5	Topic 1	6	Topic 1	7	Topic 1	8	Topic 1	6	Topic 1	9	Topic 1	7	Topic 1	8	Topic 1	5	Topic 1	8
Topic 2	7	Topic 2	7	Topic 2	7	Topic 2	4	Topic 2	11	Topic 2	10	Topic 2	8	Topic 2	8	Topic 2	10	Topic 2	7
Topic 3	8	Topic 3	7	Topic 3	6	Topic 3	8	Topic 3	3	Topic 3	5	Topic 3	5	Topic 3	4	Topic 3	5	Topic 3	5

Table 34 - Experiment #3A - Item Selection - Attempts 11 - 20

Attempt 21		Attempt 22		Attempt 23		Attempt 24		Attempt 25		Attempt 26		Attempt 27		Attempt 28		Attempt 29		Attempt 30	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 H7	42.14	1.0 H5	47.86	1.0 E6	82.86	1.0 M1	48.57	1.0 M14	58.57	1.0 M10	48.57	1.0 E7	72.86	1.0 E1	77.14	1.0 E1	77.14	1.0 M15	58.57
1.0 M11	57.86	1.0 H7	42.14	1.0 M12	65.00	1.0 M17	60.71	1.0 M18	90.71	1.0 M17	60.71	1.0 H5	47.86	1.0 M12	65.00	1.0 H5	47.86	1.0 M16	65.00
1.0 M19	50.00	1.0 M11	57.86	1.0 M14	53.57	1.0 M25	54.29	1.0 M20	59.29	1.0 M27	53.57	1.0 H4	40.00	1.0 M17	60.71	1.0 H5	47.86	1.0 M27	53.57
1.0 M27	53.57	1.0 M14	53.57	1.0 M16	65.00	1.0 M27	53.57	2.0 E11	82.14	1.0 M8	55.00	1.0 M18	50.71	1.0 M22	61.43	1.0 H6	52.86	1.0 M5	57.14
1.0 M28	55.71	1.0 M17	60.71	1.0 M17	60.71	2.0 E12	74.29	2.0 E15	75.71	2.0 E1	75.00	1.0 M22	61.43	1.0 M28	52.86	1.0 M18	50.71	2.0 E16	75.71
1.0 M8	55.00	1.0 M18	50.71	1.0 M22	61.43	2.0 E14	75.00	2.0 E16	75.71	2.0 E12	74.29	1.0 M4	62.86	1.0 M3	57.14	1.0 M28	55.71	2.0 E2	72.86
2.0 E1	75.00	1.0 M20	59.29	1.0 M3	57.14	2.0 E16	75.71	2.0 E2	72.86	2.0 E13	77.86	1.0 M9	57.14	1.0 M7	63.57	1.0 M3	57.14	3.0 M10	64.29
2.0 E10	84.29	1.0 M26	52.86	2.0 E1	75.00	2.0 E22	74.29	2.0 E21	75.71	2.0 E15	75.71	2.0 E19	75.00	2.0 E1	75.00	2.0 E21	73.57	2.0 M11	57.14
2.0 E23	76.43	2.0 E16	75.71	2.0 E13	77.86	2.0 E3	73.57	2.0 E4	77.14	2.0 E25	76.43	2.0 E20	75.71	2.0 E15	75.71	2.0 M1	60.00	2.0 M13	64.29
2.0 E8	73.57	2.0 E4	77.14	2.0 E6	75.57	2.0 E5	74.29	2.0 M15	61.43	2.0 E24	72.14	2.0 E22	74.29	2.0 E7	80.00	2.0 M11	57.14	2.0 M15	61.43
2.0 E8	81.43	2.0 M25	60.00	2.0 E8	81.43	2.0 E8	81.43	2.0 M17	67.86	2.0 E7	80.00	2.0 E5	74.29	2.0 M19	60.71	2.0 M15	61.43	2.0 M17	67.86
2.0 M1	60.00	2.0 M7	65.00	3.0 E13	72.86	2.0 M10	64.29	2.0 M21	59.29	2.0 E9	77.14	2.0 M1	60.00	2.0 M23	57.86	2.0 M16	64.29	2.0 M18	59.29
2.0 M21	59.29	2.0 M9	65.00	3.0 E6	79.29	2.0 M18	59.29	2.0 M22	58.57	1.0 M10	64.29	2.0 M25	57.86	2.0 M26	57.14	2.0 M25	60.00	3.0 E10	78.57
2.0 M23	57.86	3.0 E5	78.57	3.0 M1	67.14	2.0 M25	60.00	2.0 M23	57.86	2.0 M11	57.14	3.0 E12	77.86	2.0 M7	65.00	2.0 M27	57.86	3.0 E14	72.14
2.0 M24	60.71	3.0 M11	57.86	3.0 M10	59.29	2.0 M9	65.00	2.0 M24	60.71	2.0 M13	64.29	3.0 E3	77.14	3.0 E12	77.86	2.0 M3	71.43	3.0 E5	78.57
2.0 M3	71.43	3.0 M12	55.00	3.0 M11	57.86	3.0 E13	72.86	2.0 M5	60.71	3.0 E10	78.57	3.0 M13	60.00	3.0 E7	79.57	2.0 M4	67.14	3.0 M1	67.14
3.0 M19	70.00	3.0 M13	60.00	3.0 M20	60.00	3.0 E3	77.14	2.0 M6	70.00	3.0 H1	45.00	3.0 M18	60.00	3.0 M11	57.86	2.0 M7	65.00	3.0 M14	60.71
3.0 M21	62.86	3.0 M18	60.00	3.0 M21	62.86	3.0 M13	60.00	3.0 E1	74.29	3.0 M13	60.00	3.0 M2	63.57	3.0 M17	60.00	3.0 E2	77.14	3.0 M19	70.00
3.0 M7	57.14	3.0 M2	63.57	3.0 M23	67.86	3.0 M19	70.00	3.0 E11	77.86	3.0 M19	70.00	3.0 M20	60.00	3.0 M20	60.00	3.0 H1	45.00	3.0 M22	57.86
3.0 M8	65.00	3.0 M20	60.00	3.0 M7	57.14	3.0 M20	60.00	3.0 E2	77.14	3.0 M21	62.86	3.0 M6	58.57	3.0 M6	58.57	3.0 M8	65.00	3.0 M4	62.86
Difficulty	63.46	Difficulty	60.14	Difficulty	66.89	Difficulty	66.72	Difficulty	67.32	Difficulty	66.43	Difficulty	63.36	Difficulty	64.86	Difficulty	59.71	Difficulty	65.25
Easy	5	Easy	3	Easy	7	Easy	9	Easy	9	Easy	9	Easy	7	Easy	6	Easy	3	Easy	5
Moderate	14	Moderate	15	Moderate	13	Moderate	11	Moderate	11	Moderate	10	Moderate	11	Moderate	14	Moderate	13	Moderate	15
Hard	1	Hard	2	Hard	0	Hard	0	Hard	0	Hard	1	Hard	2	Hard	0	Hard	4	Hard	0
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	6	Topic 1	8	Topic 1	7	Topic 1	4	Topic 1	3	Topic 1	4	Topic 1	7	Topic 1	7	Topic 1	7	Topic 1	4
Topic 2	10	Topic 2	5	Topic 2	4	Topic 2	11	Topic 2	14	Topic 2	11	Topic 2	8	Topic 2	7	Topic 2	10	Topic 2	8
Topic 3	4	Topic 3	7	Topic 3	9	Topic 3	5	Topic 3	3	Topic 3	5	Topic 3	7	Topic 3	6	Topic 3	9	Topic 3	8

Table 35 - Experiment #3A - Item Selection - Attempts 21 - 30

The item selection difficulty varied widely with each iteration (see ‘Difficulty’ columns in tables 33, 34, and 35). The target cut score/difficulty was 64.37. The randomization produced a difficulty range between 59.71 and 67.54 with average of 64.15. The standard deviation of the scores was 2.44 with a 95% confidence interval of 0.878 which means that the true population mean is between 63.27 and 65.03 of the 30 samples. The kurtosis of the average difficulty is -0.859 and the skewness is -0.403. The number of items at each difficulty level from each topic varied with each iteration. Table 36 provides a summary of the statistics for the sample. Figure 10 illustrates the standard distribution curve of the sample.

Sample Difficulty Statistics	
Target Cut Score	64.37
Mean difficulty	64.15
Median	64.41
Minimum	59.71
Maximum	67.54
Variance Target vs. Mean	0.02
Standard Deviation all Averages	2.44
95% Confidence Score	0.873784614
Kurtosis	-0.858833055
Skewness	-0.403006617

Table 36 - Difficulty Statistics for Experiment #3A

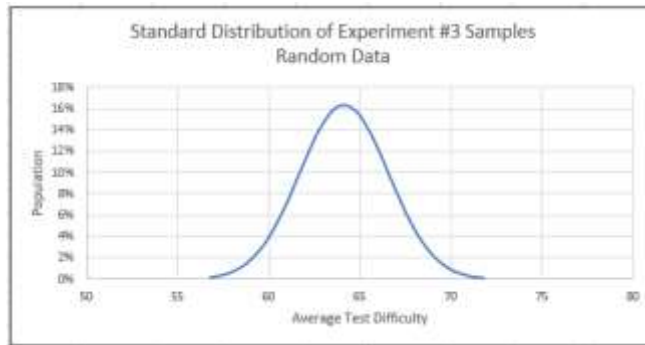


Figure 10 - Standard Distribution of Test Difficulty - Experiment #3A - Random Selection

The content (topic) coverage was erratic as illustrated by the three-color (delineated by sections) display and 'Total From Topic' columns in tables 33, 34, and 35.

Conclusion: All topics were not covered equally in either difficulty or content. Although the average (mean) difficulty of the sample is within an acceptable range of the desired difficulty/cut score the standard deviation indicates a significant variation among attempts.

Experiment #3B – Stratified Randomization – Real Client #2 Data

Using the same data from experiment #3A, following the recommended test design, the assessment design function of Questionmark OnDemand was instructed to select 20 items in a stratified random fashion from the sub-topics, following the recommended design shown in table 32 (figure 11).

Question selections
1 random question(s) from topic 'FAIRNESS RESEARCH 3/1.0 TOPIC 1/1.0 HARD' excluding subtopics (Avoid previously delivered)
4 random question(s) from topic 'FAIRNESS RESEARCH 3/1.0 TOPIC 1/1.0 MODERATE' excluding subtopics (Avoid previously delivered)
1 random question(s) from topic 'FAIRNESS RESEARCH 3/1.0 TOPIC 1/1.0 EASY' excluding subtopics (Avoid previously delivered)
1 random question(s) from topic 'FAIRNESS RESEARCH 3/2.0 TOPIC 2/2.0 HARD' excluding subtopics (Avoid previously delivered)
4 random question(s) from topic 'FAIRNESS RESEARCH 3/2.0 TOPIC 2/2.0 MODERATE' excluding subtopics (Avoid previously delivered)
3 random question(s) from topic 'FAIRNESS RESEARCH 3/2.0 TOPIC 2/2.0 EASY' excluding subtopics (Avoid previously delivered)
1 random question(s) from topic 'FAIRNESS RESEARCH 3/3.0 TOPIC 3/3.0 HARD' excluding subtopics (Avoid previously delivered)
3 random question(s) from topic 'FAIRNESS RESEARCH 3/3.0 TOPIC 3/3.0 MODERATE' excluding subtopics (Avoid previously delivered)
2 random question(s) from topic 'FAIRNESS RESEARCH 3/3.0 TOPIC 3/3.0 EASY' excluding subtopics (Avoid previously delivered)

Figure 11 - Item Selection Criteria - Experiment #3B

Thirty (n=30) iterations of a 20-question test were generated as illustrated in tables 37, 38, and 39.

Experiment 3B - Directed Random Selection of 20 items from all 3 topics. Real Client Data Desired target difficulty is 64.37																			
Attempt 1		Attempt 2		Attempt 3		Attempt 4		Attempt 5		Attempt 6		Attempt 7		Attempt 8		Attempt 9		Attempt 10	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 H4	40.00	1.0 H3	47.86	1.0 H6	32.86	1.0 H2	44.29	1.0 H3	47.86	1.0 H7	42.14	1.0 H3	47.86	1.0 H7	42.14	1.0 H2	44.29	1.0 H2	44.29
1.0 M7	63.57	1.0 M28	55.71	1.0 M18	50.71	1.0 M25	54.29	1.0 M16	65.00	1.0 M8	55.00	1.0 M6	62.14	1.0 M27	53.57	1.0 M20	59.29	1.0 M6	62.14
1.0 M20	59.29	1.0 M11	57.86	1.0 M23	71.43	1.0 M28	55.71	1.0 M10	48.57	1.0 M26	52.86	1.0 M5	57.14	1.0 M10	48.57	1.0 M14	53.57	1.0 M17	60.71
1.0 M28	55.71	1.0 M13	56.43	1.0 M1	48.57	1.0 M14	53.57	1.0 M26	52.86	1.0 M13	56.43	1.0 M4	62.86	1.0 M14	53.57	1.0 M8	55.00	1.0 M27	53.57
1.0 M23	61.43	1.0 M14	53.57	1.0 M3	57.14	1.0 M17	60.71	1.0 M10	59.86	1.0 M4	62.86	1.0 M20	59.29	1.0 M15	56.43	1.0 M23	71.43	1.0 M21	57.86
1.0 E5	72.86	1.0 E2	75.00	1.0 E2	75.00	1.0 E1	77.14	1.0 E7	72.86	1.0 E7	72.86	1.0 E5	72.86	1.0 E7	72.86	1.0 E6	82.86	1.0 E6	82.86
2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43
2.0 M8	67.86	2.0 M12	60.71	2.0 M18	59.29	2.0 M26	57.14	2.0 M11	57.14	2.0 M1	60.00	2.0 M24	80.71	2.0 M25	60.00	2.0 M17	67.86	2.0 M2	69.29
2.0 M21	59.29	2.0 M21	59.29	2.0 M13	64.29	2.0 M4	67.14	2.0 M12	60.71	2.0 M5	60.71	2.0 M2	69.29	2.0 M17	67.86	2.0 M27	57.86	2.0 M10	64.29
2.0 M26	57.14	2.0 M26	57.14	2.0 M19	60.71	2.0 M15	61.43	2.0 M25	60.00	2.0 M20	67.14	2.0 M14	67.86	2.0 M8	67.86	2.0 M1	60.00	2.0 M12	60.71
2.0 M25	60.00	2.0 M18	59.29	2.0 M9	65.00	2.0 M9	65.00	2.0 M24	80.71	2.0 M20	60.00	2.0 M15	61.43	2.0 M4	67.14	2.0 M10	64.29	2.0 M26	57.14
2.0 E20	73.71	2.0 E15	75.71	2.0 E3	73.57	2.0 E23	76.43	2.0 E3	73.57	2.0 E8	81.43	2.0 E19	75.00	2.0 E24	72.14	2.0 E20	75.71	2.0 E6	73.57
2.0 E10	84.29	2.0 E16	75.71	2.0 E11	82.14	2.0 E15	75.71	2.0 E19	75.00	2.0 E21	73.57	2.0 E5	74.29	2.0 E17	77.14	2.0 E11	82.14	2.0 E4	77.14
2.0 E5	74.29	2.0 E9	77.14	2.0 E13	77.86	2.0 E2	72.86	2.0 E14	75.00	2.0 E1	74.29	2.0 E18	77.86	2.0 E16	75.71	2.0 E23	74.29	2.0 E9	77.14
3.0 H2	46.43	3.0 H1	45.00	3.0 H2	46.43	3.0 H1	45.00	3.0 H2	46.43	3.0 H2	46.43	3.0 H1	45.00	3.0 H1	45.00	3.0 H2	46.43	3.0 H1	45.00
3.0 M23	67.86	3.0 M24	54.29	3.0 M16	68.57	3.0 M10	59.29	3.0 M7	57.14	3.0 M4	62.86	3.0 M6	58.57	3.0 M19	70.00	3.0 M22	57.86	3.0 M13	60.00
3.0 M2	63.57	3.0 M18	60.00	3.0 M13	60.00	3.0 M14	60.71	3.0 M12	55.00	3.0 M17	60.00	3.0 M16	68.57	3.0 M14	60.71	3.0 M7	57.14	3.0 M23	67.86
3.0 M6	58.57	3.0 M21	62.86	3.0 M21	62.86	3.0 M1	67.14	3.0 M21	62.86	3.0 M8	65.00	3.0 M8	65.00	3.0 M5	71.43	3.0 M17	60.00	3.0 M9	56.43
3.0 E7	73.57	3.0 E7	73.57	3.0 E9	77.14	3.0 E2	77.14	3.0 E3	77.14	3.0 E6	79.29	3.0 E3	77.14	3.0 E4	78.57	3.0 E14	72.14	3.0 E6	79.29
3.0 E2	77.14	3.0 E6	79.29	3.0 E4	78.57	3.0 E12	77.86	3.0 E2	77.14	3.0 E4	78.57	3.0 E7	75.57	3.0 E11	77.86	3.0 E6	79.29	3.0 E10	78.57
Difficulty	63.25	Difficulty	61.64	Difficulty	62.93	Difficulty	62.75	Difficulty	61.56	Difficulty	62.89	Difficulty	64.14	Difficulty	63.25	Difficulty	63.39	Difficulty	63.71
Easy	6	Easy	6	Easy	6	Easy	6	Easy	6	Easy	6	Easy	6	Easy	6	Easy	6	Easy	6
Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11
Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6
Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8
Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6

Table 37 - Experiment #3B Item Selections - Attempts 1 - 10

Attempt 11		Attempt 12		Attempt 13		Attempt 14		Attempt 15		Attempt 16		Attempt 17		Attempt 18		Attempt 19		Attempt 20	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 H6	32.86	1.0 H4	40.00	1.0 H5	47.86	1.0 H5	47.86	1.0 H5	47.86	1.0 H1	41.43	1.0 H7	42.14	1.0 H3	47.86	1.0 H1	41.43	1.0 H3	47.86
1.0 M17	60.71	1.0 M5	60.00	1.0 M14	53.57	1.0 M20	59.29	1.0 M4	62.86	1.0 M11	57.86	1.0 M21	57.86	1.0 M20	59.29	1.0 M12	65.00	1.0 M10	48.57
1.0 M19	50.00	1.0 M11	57.86	1.0 M17	60.71	1.0 M26	52.86	1.0 M8	55.00	1.0 M20	59.29	1.0 M22	61.43	1.0 M22	61.43	1.0 M14	53.57	1.0 M28	55.71
1.0 M1	48.57	1.0 M19	50.00	1.0 M7	63.57	1.0 M6	62.14	1.0 M5	60.00	1.0 M18	65.00	1.0 M9	57.14	1.0 M18	50.71	1.0 M25	54.29	1.0 M20	59.29
1.0 M23	71.43	1.0 M24	56.43	1.0 M25	71.43	1.0 M24	56.43	1.0 M20	59.29	1.0 M17	60.71	1.0 M11	57.86	1.0 M4	62.86	1.0 M20	59.29	1.0 M5	60.00
1.0 E3	77.14	1.0 E5	72.86	1.0 E5	72.86	1.0 E1	77.14	1.0 E4	84.29	1.0 E6	82.86	1.0 E5	72.86	1.0 E2	75.00	1.0 E7	72.86	1.0 E7	72.86
2.0 H1	46.43	2.0 H1	46.43	2.0 H2	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43
2.0 M17	67.86	2.0 M2	69.29	2.0 M24	60.71	2.0 M26	57.14	2.0 M3	71.43	2.0 M27	57.86	2.0 M24	60.71	2.0 M2	69.29	2.0 M3	71.43	2.0 M25	60.00
2.0 M10	64.29	2.0 M12	60.71	2.0 M12	60.71	2.0 M6	70.00	2.0 M24	80.71	2.0 M12	60.71	2.0 M20	67.14	2.0 M23	57.86	2.0 M8	70.00	2.0 M16	64.29
2.0 M6	70.00	2.0 M10	64.29	2.0 M11	57.14	2.0 M10	64.29	2.0 M17	67.86	2.0 M4	67.14	2.0 M22	58.57	2.0 M18	59.29	2.0 M5	60.71	2.0 M1	60.00
2.0 M26	57.14	2.0 M16	64.29	2.0 M4	67.14	2.0 M12	60.71	2.0 M12	60.71	2.0 M6	70.00	2.0 M9	65.00	2.0 M6	70.00	2.0 M25	57.14	2.0 M22	58.57
2.0 E1	75.00	2.0 E20	75.71	2.0 E5	74.29	2.0 E3	74.29	2.0 E4	77.14	2.0 E5	74.29	2.0 E24	72.14	2.0 E21	73.57	2.0 E4	77.14	2.0 E8	81.43
2.0 E19	75.00	2.0 E15	75.71	2.0 E21	73.57	2.0 E11	82.14	2.0 E5	74.29	2.0 E22	74.29	2.0 E21	73.57	2.0 E14	75.00	2.0 E11	82.14	2.0 E7	80.00
2.0 E13	77.86	2.0 E23	76.43	2.0 E19	75.00	2.0 E3	73.57	2.0 E6	73.57	2.0 E16	75.71	2.0 E7	80.00	2.0 E9	77.14	2.0 E12	74.29	2.0 E21	73.57
3.0 H2	46.43	3.0 H1	45.00	3.0 H2	46.43	3.0 H2	46.43	3.0 H2	46.43	3.0 H2	46.43	3.0 H1	45.00	3.0 H1	45.00	3.0 H1	45.00	3.0 H1	45.00
3.0 M8	65.00	3.0 M8	65.00	3.0 M21	62.86	3.0 M2	63.57	3.0 M14	60.71	3.0 M22	57.86	3.0 M12	55.00	3.0 M16	68.57	3.0 M13	60.00	3.0 M1	67.14
3.0 M5	71.43	3.0 M21	62.86	3.0 M8	65.00	3.0 M9	56.43	3.0 M4	62.86	3.0 M15	49.29	3.0 M5	71.43	3.0 M4	62.86	3.0 M14	60.71	3.0 M14	60.71
3.0 M14	60.71	3.0 M19	70.00	3.0 M3	61.43	3.0 M12	55.00	3.0 M23	67.86	3.0 M4	62.86	3.0 M15	49.29	3.0 M11	57.86	3.0 M11	57.86	3.0 M17	60.00
3.0 E6	79.29	3.0 E14	72.14	3.0 E13	72.86	3.0 E9	77.14	3.0 E4	78.57	3.0 E2	77.14	3.0 E6	79.29	3.0 E10	78.57	3.0 E3	77.14	3.0 E1	74.29
3.0 E14	72.14	3.0 E11	77.86	3.0 E8	73.57	3.0 E13	72.86	3.0 E7	73.57	3.0 E5	78.57	3.0 E3	77.14	3.0 E9	77.14	3.0 E1	74.29	3.0 E6	79.29
Difficulty	63.46	Difficulty	63.14	Difficulty	63.36	Difficulty	62.79	Difficulty	64.57	Difficulty	63.29	Difficulty	62.50	Difficulty	63.79	Difficulty	63.04	Difficulty	62.75
Easy	8	Easy	8	Easy	6	Easy	6	Easy	6	Easy	6	Easy	8	Easy	6	Easy	6	Easy	6
Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11
Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6
Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8
Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6

Table 38 - Experiment #3B Item Selections - Attempts 11 - 20

Attempt 21		Attempt 22		Attempt 23		Attempt 24		Attempt 25		Attempt 26		Attempt 27		Attempt 28		Attempt 29		Attempt 30	
QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE	QID	SCORE
1.0 H3	47.86	1.0 H6	32.86	1.0 H4	40.00	1.0 H1	41.43	1.0 H7	42.14	1.0 H5	47.86	1.0 H3	47.86	1.0 H7	42.14	1.0 H4	40.00	1.0 H5	47.86
1.0 M15	58.57	1.0 M4	62.86	1.0 M24	56.43	1.0 M5	60.00	1.0 M1	48.57	1.0 M21	57.86	1.0 M19	50.00	1.0 M27	53.57	1.0 M8	57.14	1.0 M16	65.00
1.0 M12	65.00	1.0 M2	62.14	1.0 M6	55.00	1.0 M12	65.00	1.0 M24	56.43	1.0 M27	53.57	1.0 M21	57.86	1.0 M11	57.86	1.0 M8	55.00	1.0 M20	59.29
1.0 M17	60.71	1.0 M22	61.43	1.0 M18	50.71	1.0 M15	58.57	1.0 M5	57.14	1.0 M6	60.00	1.0 M6	62.14	1.0 M5	60.00	1.0 M12	65.00	1.0 M1	48.57
1.0 M2	62.14	1.0 M19	50.00	1.0 M2	62.14	1.0 M19	50.00	1.0 M9	57.14	1.0 M15	58.57	1.0 M23	71.43	1.0 M24	56.43	1.0 M19	50.00	1.0 M27	53.57
1.0 E4	84.29	1.0 E1	77.14	1.0 E7	72.86	1.0 E6	82.86	1.0 E6	82.86	1.0 E7	72.86	1.0 E4	84.29	1.0 E2	75.00	1.0 E2	75.00	1.0 E7	72.86
2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43	2.0 H1	46.43
2.0 M9	65.00	2.0 M16	64.29	2.0 M9	65.00	2.0 M1	60.00	2.0 M25	60.00	2.0 M27	57.86	2.0 M17	67.86	2.0 M5	60.71	2.0 M22	58.57	2.0 M10	64.29
2.0 M27	57.86	2.0 M27	57.86	2.0 M20	67.14	2.0 M8	67.86	2.0 M5	60.71	2.0 M14	67.86	2.0 M22	58.57	2.0 M24	60.71	2.0 M1	60.00	2.0 M25	60.00
2.0 M1	60.00	2.0 M14	67.86	2.0 M2	69.29	2.0 M11	57.14	2.0 M16	64.29	2.0 M25	60.00	2.0 M5	60.71	2.0 M7	65.00	2.0 M14	67.86	2.0 M23	57.86
2.0 M6	70.00	2.0 M13	64.29	2.0 M5	60.71	2.0 M19	60.71	2.0 M15	61.43	2.0 M7	65.00	2.0 M24	60.71	2.0 M10	64.29	2.0 M5	60.71	2.0 M22	58.57
2.0 E10	84.29	2.0 E15	75.71	2.0 E23	76.43	2.0 E9	77.14	2.0 E8	81.43	2.0 E9	81.43	2.0 E21	73.57	2.0 E19	75.00	2.0 E11	82.14	2.0 E15	75.71
2.0 E3	73.57	2.0 E1	75.00	2.0 E14	75.00	2.0 E2	73.86	2.0 E14	75.00	2.0 E19	75.00	2.0 E10	75.71	2.0 E14	72.14	2.0 E9	77.14	2.0 E9	77.14
2.0 E8	81.43	2.0 E24	72.14	2.0 E12	74.29	2.0 E13	77.86	2.0 E15	75.71	2.0 E5	74.29	2.0 E10	84.29	2.0 E21	73.57	2.0 E8	81.43	2.0 E21	73.57
3.0 H1	45.00	3.0 H1	45.00	3.0 H1	45.00	3.0 H1	45.00	3.0 H2	46.43	3.0 H1	45.00	3.0 H2	46.43	3.0 H2	46.43	3.0 H2	46.43	3.0 H2	46.43
3.0 M14	60.71	3.0 M6	58.57	3.0 M11	57.86	3.0 M23	67.86	3.0 M1	67.14	3.0 M11	57.86	3.0 M2	63.57	3.0 M11	57.86	3.0 M5	61.43	3.0 M17	60.00
3.0 M9	56.43	3.0 M5	71.43	3.0 M5	61.43	3.0 M22	57.86	3.0 M2	63.57	3.0 M8	65.00	3.0 M5	61.43	3.0 M10	59.29	3.0 M10	59.29	3.0 M7	57.14
3.0 M11	57.86	3.0 M17	60.00	3.0 M24	54.29	3.0 M9	56.43	3.0 M6	65.00	3.0 M23	67.86	3.0 M24	54.29	3.0 M9	56.43	3.0 M4	62.86	3.0 M18	60.00
3.0 E4	78.57	3.0 E9	77.14	3.0 E11	77.86	3.0 E14	72.14	3.0 E10	78.57	3.0 E12	77.86	3.0 E1	74.29	3.0 E9	77.14	3.0 E9	77.14	3.0 E2	77.14
3.0 E6	79.29	3.0 E3	77.14	3.0 E1	74.29	3.0 E7	73.57	3.0 E9	77.14	3.0 E9	77.14	3.0 E2	77.14	3.0 E4	78.57	3.0 E2	77.14	3.0 E3	77.14
Difficulty	64.75	Difficulty	62.96	Difficulty	62.11	Difficulty	62.54	Difficulty	63.36	Difficulty	63.47	Difficulty	63.93	Difficulty	61.93	Difficulty	63.04	Difficulty	61.93
Easy	6	Easy	6	Easy	6	Easy	6	Easy	6	Easy	6	Easy	6	Easy	6	Easy	6	Easy	6
Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11	Moderate	11
Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3	Hard	3
Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic		Total From Topic	
Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6	Topic 1	6
Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8	Topic 2	8
Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6	Topic 3	6

Table 39 - Experiment #3B Item Selections - Attempts 21 - 30

The item selection difficulty remained constant as stratified with each iteration (see ‘Difficulty’ columns in tables 37, 38, and 39). The target cut score/difficulty was 64.37. The stratified randomization produced a difficulty range between 61.93 and 64.75 with average (mean) of 63.13. The standard deviation of the scores was 0.76 with a 95% confidence interval of 0.274 which means that the true population mean is between 62.86 and 63.40 of the 30 samples. The kurtosis of the average difficulty is 0.127 and the skewness is 0.351. The number of items at each difficulty level from each topic varied with each iteration. Table 28 provides a summary of the statistics for the sample. Figure 12 illustrates the standard distribution curve of the sample.

Sample Difficulty Statistics	
Target Cut Score	64.37
Mean difficulty	63.13
Median	63.09
Minimum	61.93
Maximum	64.75
Variance Target vs. Mean	0.76
Standard Deviation all Averages	0.76
95% Confidence Score	0.273697889
Kurtosis	0.127180112
Skewness	0.351166789

Table 40 - Difficulty Statistics for Experiment #3B

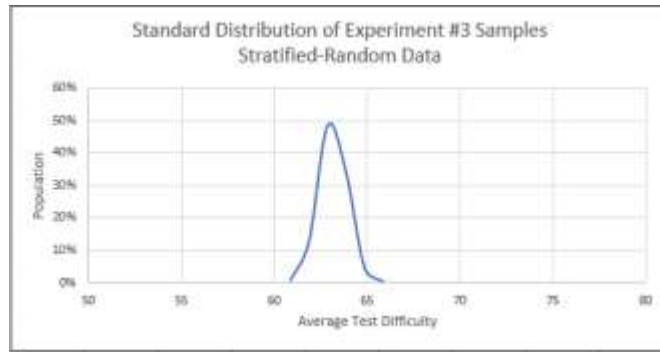


Figure 12 - Standard Distribution of Test Difficulty - Experiment #3B - Stratified Random Selection

The content (topic) coverage was equal as stratified for each iteration as illustrated by the four-color (delineated by sections) display and ‘Total From Topic’ columns in tables 37, 38, and 39.

Conclusion: All topics were covered equally as desired in both difficulty and content. Comparing the spread of test difficulty scores in figures 10 and 12 shows that the stratified randomization consistently produced tests well within an acceptable range to meet the desired cut score of 64.37. Although the simple randomization in experiment 3A produced tests with an average (mean) difficulty (64.15) closer to the desired difficulty (64.37) than the stratified randomization in experiment 3B (63.13), the standard deviation of the results in experiment 3B (0.76) indicated significantly less variance from attempt to attempt than the standard deviation produced by experiment 3A (2.44).

CONCLUSIONS

Key factors in maintaining defensibility and fairness of assessments are being able to justify the cut or passing score as well as maintain validity. In that regard, each iteration of a parallel assessment or test must be parallel in both difficulty and topic coverage otherwise there is unfair bias each time an assessment is generated. Conclusions drawn from the experiments described in this paper are as follow:

1. If test items are selected at random from a test-item database, without regard to either difficulty or subject matter, the resulting assessments will be inconsistent in coverage of topics and difficulty. This is evidenced by the results of experiments 1A, 2A, and 3A.
 - a. Selection of the number of test-items from each topic was inconsistent with each iteration of the randomly generated assessments.
 - i. Experiment 1A tables 6, 7, and 8
 - ii. Experiment 2A tables 20, 21, and 22
 - iii. Experiment 3A tables 33, 34, and 35
 - b. Each item in the test-item database was empirically assigned a difficulty score ranging from .25 (25%) to .95 (95%) using the Parry Method (a variation of the Angoff Method). Each item was then classified as Easy, Moderate, or Hard based upon the score range determined by dividing the entire score range into thirds (see table A-1) The “cut” score for the entire database was determined by averaging the results of item score. This cut score was the target for each iteration of the assessment.

- i. The mean (average) difficulty or cut score for each experiment of randomly generated assessments was within one point of the target cut score, and actually closer to the desired cut score than the mean (average) using stratified-random selection (table 41).
 - ii. The standard deviation among each iteration was higher than those generated using stratified-random selection which translates to a wider spread or variance of actual difficulties among iterations. This is evidenced by the comparison of the standard score distribution plots (Figure 13).
- 2. If test items are selected from a test-item database, using stratified-random selection by difficulty and subject matter, the resulting assessments will be consistent in coverage of topics and difficulty. This is evidenced by the results of experiments 1B, 2B, and 3B.
 - a. Selection of the number of test-items from each topic was consistent with each iteration of the randomly generated assessments. The selection of the number of items from each topic was a forced selection based upon the number of items available in each topic vs. the desired test length. The philosophy behind this selection is described in Appendix A.
 - i. Experiment 1B tables 6, 7, and 8
 - ii. Experiment 2B tables 20, 21, and 22
 - iii. Experiment 3B tables 33, 34, and 35
 - b. As previously stated, each item in the test-item database was empirically assigned a difficulty score and placed into appropriate topic and difficulty folders.
 - i. Stratified-random selection, based upon difficulty as well as topic, consistently produced assessments within several points of the target cut score. The assessment difficulty in relation to the target cut score varied slightly with each iteration.
 - 1. Experiment 1B tables 6, 7, and 8
 - 2. Experiment 2B tables 20, 21, and 22
 - 3. Experiment 3B tables 33, 34, and 35

- ii. Although the difference between the mean (average) difficulty or cut score and the target cut score for each experiment of stratified-random generated assessments was slightly more than that of the randomly generated assessments, (table 41), the standard deviation among iterations was lower which translates to a smaller spread or variance of actual difficulties among iterations. This is evidenced by the comparison of the standard score distribution plots (Figure 13).

Target Difficulty (Cut-Score) vs. Actual Results			
Experiment #	Target	Random Selection	Stratified-Random Selection
1	58.21	58.54	58.84
2	76.32	75.87	74.11
3	64.73	64.15	63.13

Table 41 - Target Difficulty (Cut-Score) vs. Actual Results

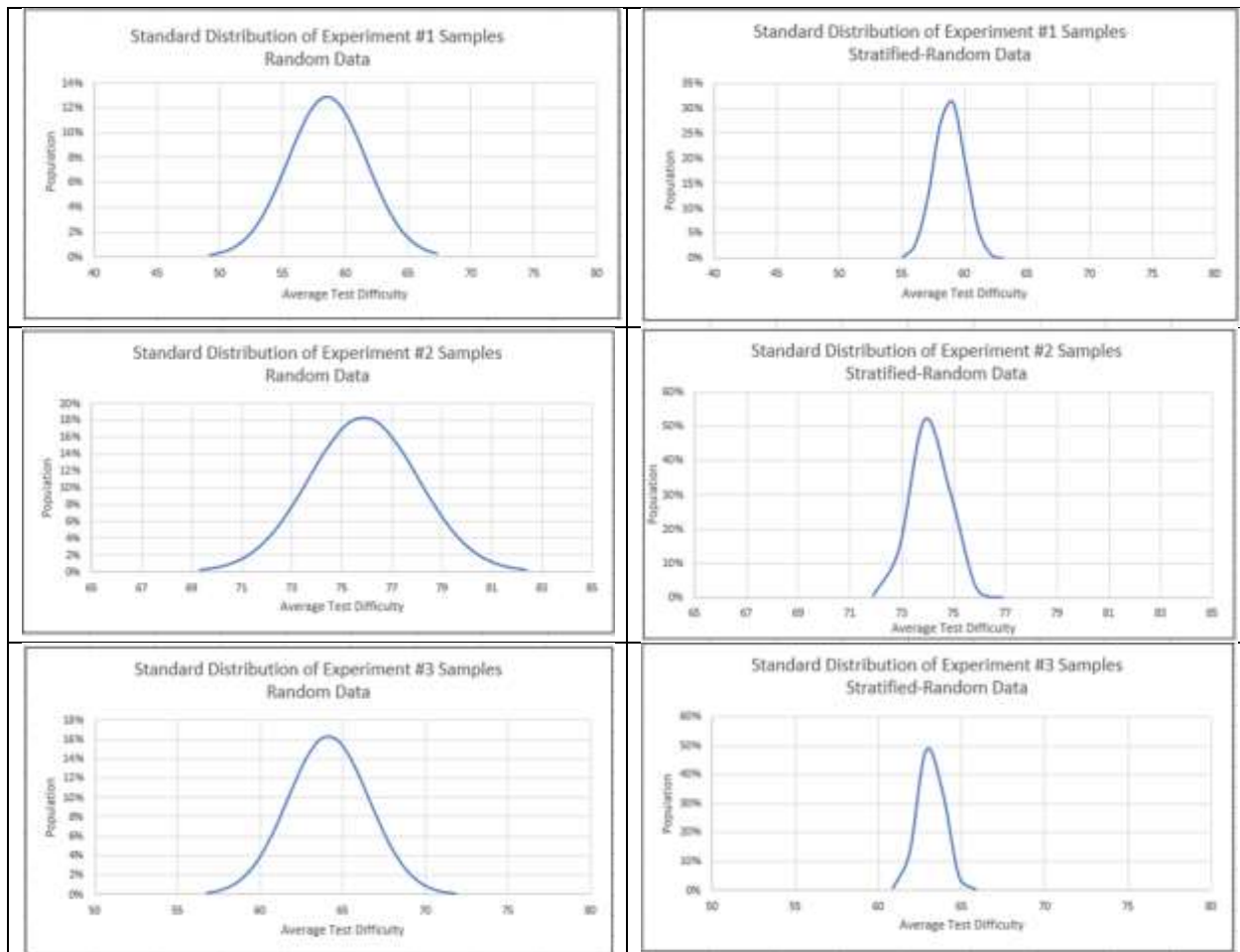


Figure 13 - Comparison of Standard Distribution of Difficulty - Random Selection vs. Stratified-Random Selection

RECOMMENDATIONS

In order to maintain fairness and ensure tests are valid, reliable, without bias and defensible when generated from a test-item database:

- Test-items must be constructed using universally recognized standards
- Cut scores should be established using a recognized test-centered method or, if appropriate, a test-taker centered method, because arbitrary methods are not defensible
- Each item in a test-item database should be evaluated by a panel of expert judges and a difficulty score or rating established based upon the agreed upon MAC level of the target test-taker
- Test items should be selected using stratified randomization based upon both topic coverage as well as item difficulty to ensure equitable parallel assessments are generated
- Tests should not be generated in a pure random fashion from a test-item database without regard to content because content coverage will be erratic
- Tests should not be generated in a pure random fashion from test-item database without regard to difficulty of test-items because difficulty among tests will be erratic

LIST OF FIGURES

Figure 1 – Designing Criterion-Referenced Tests.....	8
Figure 2 - Questionmark OnDemand Item Selection Criteria for Hypothetical Data	18
Figure 3 - Standard Distribution of Test Difficulty - Experiment #1A - Random Selection	20
Figure 4 - Topic, Sub-topic Structure Used in Experiments	21
Figure 5 - Item Selection Criteria - Experiment #1B.....	22
Figure 6 - Standard Distribution of Test Difficulty - Experiment #1B - Stratified Random Selection	24
Figure 7 - Standard Distribution of Test Difficulty - Experiment #2A - Random Selection	29
Figure 8 - Item Selection Criteria - Experiment #2B.....	30
Figure 9 - Standard Distribution of Test Difficulty - Experiment #2B - Stratified Random Selection	32
Figure 10 - Standard Distribution of Test Difficulty - Experiment #3A - Random Selection	37
Figure 11 - Item Selection Criteria - Experiment #3B.....	37
Figure 12 - Standard Distribution of Test Difficulty - Experiment #3B - Stratified Random Selection	40
Figure 13 - Comparison of Standard Distribution of Difficulty - Random Selection vs. Stratified-Random Selection.....	42

LIST OF TABLES

Table 1 - Number of Test-Items per Objective	9
Table 2 - Topic Weights for Electrical Safety Assessment	10
Table 3 - Test Time Tool	11
Table 4 - Experiment # 1- Cut Score Calculation Tool (partial view) showing hypothetical data.....	17
Table 5 – Experiment #1 - Recommended Test Difficulty Distribution for Hypothetical Data – Random Selection.....	17
Table 6 - Experiment #1A - Item Selection - Attempts 1 - 10	18
Table 7 - Experiment #1A - Item Selection - Attempts 11 - 20	19
Table 8 - Experiment #1A - Item Selection - Attempts 21 - 30	19
Table 9 - Difficulty Statistics for Experiment #1A	20
Table 10 - Recommended Test Design - Stratified Randomization - Experiment #1B.....	21
Table 11 - Experiment #1B Item Selections - Attempts 1 - 10	22
Table 12 - Experiment #1B Item Selections - Attempts 11 - 20	23
Table 13 - Experiment #1B Item Selections - Attempts 21 - 30	23
Table 14 - Difficulty Statistics for Experiment #1B.....	24
Table 15 - Experiment #2A - Difficulty Calculations - Real Data - Topic 1.....	25
Table 16 - Experiment #2A - Difficulty Calculations - Real Data - Topic 2.....	25
Table 17 - Experiment #2A - Difficulty Calculations - Real Data - Topic 3.....	26
Table 18 - Experiment #2 - Item Difficulty Distribution by Topic.....	26
Table 19 - Experiment #2 - Recommended Test Design	26
Table 20 - Experiment #2A - Item Selection - Attempts 1 - 10	27
Table 21 - Experiment #2A - Item Selection - Attempts 11 - 20	27
Table 22 - Experiment #2A - Item Selection - Attempts 21 - 30	28
Table 23 - Difficulty Statistics for Experiment #2A	28
Table 24 - Experiment #2B Item Selections - Attempts 1 - 10	30
Table 25 - Experiment #2B Item Selections - Attempts 11 - 20	31
Table 26 - Experiment #2B Item Selections - Attempts 21 - 30	31
Table 27 - Difficulty Statistics for Experiment #2B.....	32
Table 28 - Experiment #3A - Difficulty Calculations - Real Data - Topic 1.....	33
Table 29 - Experiment #3A - Difficulty Calculations - Real Data - Topic 2.....	33
Table 30 - Experiment #3A - Difficulty Calculations - Real Data - Topic 3.....	34
Table 31 - Experiment #3 - Item Difficulty Distribution by Topic.....	34
Table 32 - Experiment #3 - Recommended Test Design	34
Table 33 - Experiment #3A - Item Selection - Attempts 1 - 10	35
Table 34 - Experiment #3A - Item Selection - Attempts 11 - 20	35
Table 35 - Experiment #3A - Item Selection - Attempts 21 - 30	36
Table 36 - Difficulty Statistics for Experiment #3A	36
Table 37 - Experiment #3B Item Selections - Attempts 1 - 10	38
Table 38 - Experiment #3B Item Selections - Attempts 11 - 20	38
Table 39 - Experiment #3B Item Selections - Attempts 21 - 30	39
Table 40 - Difficulty Statistics for Experiment #3B.....	39
Table 41 - Target Difficulty (Cut-Score) vs. Actual Results.....	42

APPENDIX A

Description of the spreadsheet tool:

The spreadsheet tool is designed to be used in conjunction with most any test-centered cut score setting method that produces a difficulty rating for individual test-items. It has been used very successfully in conjunction with the Modified Angoff method²⁵ and is designed primarily for 4-alternative multiple-choice²⁶ test-items with allowable rating scores between .25 and .95. A version has also been developed for use with 2 through 7-alternative items with allowable rating scores adjusted to suit the number of alternatives with a minimum rating score of 14.2 for a 7-alternative item (table A-1). There are two trains of thought for the maximum rating ceiling; 1.00 or .95 (100% or 95%). The original Angoff Method allowed any value from .25 through 1.00 and only asked the judges to consider one individual. I chose a variation of the Modified Angoff Method (Parry Method) which asks the judges to consider 100 test-takers who are considered to be minimally competent (in the subject being tested) and restricts the judges to preset values allowed for simplicity. (.25, .30, .35, .40, .45, .50, .55, .60, .65, .70, .75, .80, .85, .90, .95). 1.0 (100%) is eliminated as an item that is responded to correctly by a minimally competent examinee 100% of the time is considered an unnecessary test-item in most cases.

Difficulty Values Based on Number of Alternatives			
ALTS	EASY	MODERATE	HARD
2	80.1 - 95	65.1 - 80	50 - 65
3	74.51 - 95	53.8 - 74.5	33 - 53.8
4	71.8 - 95	48.4 - 71.7	25 - 48.3
5	70.2 - 95	45.1 - 70.1	20 - 45
6	69.1 - 95	42.9 - 69	16.7 - 42.8
7	68.4 - 95	41.3 - 68.3	14.2 - 41.2

Table A- 1 - Difficulty Values Based on Number of Alternatives

The description of the spreadsheet function will use examples from the 4-alternative tool. Both tools function the same with the exception of the 2 through 7-alternative tool requires an entry by the user to identify the number of alternatives to ensure proper difficulty calculations. The tool has not been optimized for items that use multiple-response²⁷ items types but does allow for the mixing of multiple-choice items with different numbers of alternatives. It is recommended that mixing test-items with varying numbers of alternatives within the database be limited and used with extreme caution as the results of stratified random selection may generate assessments outside of the desired difficulty range or present an unbalanced mix of items with varying numbers of alternatives.

²⁵ A well-established method involving three basic steps: conceptualizing the borderline examinee, identifying specific test-items, and using expert judges to estimate what percentage of borderline examinees should respond correctly.

²⁶ A multiple-choice test-item only allows the test-taker to select one alternative as the correct/incorrect choice

²⁷ A multiple-response test-item allows the test-taker to select more than one alternative as correct/incorrect responses.

Design philosophy of the Compass Consultants, LLC spreadsheet tool: The tool is designed to assist in setting a cut score for an assessment based on the results of a test-centered cut-score rating session. Additionally, it will determine the number of items from each section at each level of difficulty (Hard, Moderate or Easy) as set by the cut-score rating of each item. This assumption is made for 4-choice, multiple choice items with a floor of 25% and a ceiling of 95%. The difficulty is then assigned based on dividing the difference between 25% and 95% by 3 to arrive at the three difficulty levels. The workbook is designed to accommodate up to ten (10) reviewers on the rating panel. The ratings assigned to each test-item by the individual judges are averaged and a difficulty score is assigned to the test-item. If the judge's individual ratings produce a standard deviation²⁸ of 10 or greater the item is flagged for discussion among the judges to either come to a consensus by modifying their ratings, retire the item as a 'bad' item or leave it as written and rated. A final test design is proposed after all items in all topics have been rated. The totals required from each section are based upon the numbers of each level of difficulty available in each section as well as the total number of items available. An assumption is made that if there are fewer items available in any particular section(s) than in other section(s), then that section is of less importance or has significantly fewer objectives. As data for each item is entered in each section, the final test design worksheet will be updated automatically.

Function of the Spreadsheet Tool


The tool is designed to accommodate 20 topics with 200 test-items per topic. The number of topics and test-items can be expanded upon request to Compass Consultants, LLC. The tool will allow for the input from up to 10 judges. Typically, 6 – 8 judges are sufficient with less than three not producing reliable ratings and more than 10 as being difficult to arrive at a consensus. Refer to table A-2 as the spreadsheet tool is described.

- As all of the heading information is entered by the facilitator, each of the individual worksheets populate automatically
- Generally, the **Enter * If New, R if Retired** block is left blank initially. As items are reviewed during the rating session, some may be omitted. If an item has NEVER been presented on an assessment and is omitted during the review, all data can be omitted and the row left blank. If additional items are added to the database in the future they should be indicated with an '*' for easy identification. It is not usually necessary to re-evaluate all items but be aware the section/topic difficulty as well as the cut score may change. If an item that is currently in use on an assessment is subsequently retired from use the worksheet row should be updated with an 'R'. This will remove any cut score/difficulty calculations from the tool and the section/topic difficulty as well as the cut score may change. To maintain defensibility for possible future challenges this data should remain in the permanent assessment documentation.
- Each test-item is given a *Question Identifier (Test-item QID)* in the test-item database before a rating session begins.
- As each expert (judge) 'score' is entered for each item the tool is updated in real time:

²⁸ Standard Deviation (σ) is a measure of how spread out numbers are.

- **Difficulty Metatag** presents the difficulty as easy, moderate, or hard and the color changes as appropriate with green indicating ‘easy’, yellow indicating ‘moderate’, and red indicating ‘hard’ based upon the judges average score
- **Average Percentage Correct (Angoff Rating)** for the item is updated to reflect the running average score
- **Standard Deviation** is updated. If the standard deviation is 10 or greater the block displays a red flag to facilitate location during the discussion phase of the rating session
- The **Topic Cut Score** block updates to reflect the current cut score/difficulty for the topic
- A running total of the number and percentage of items at each difficulty level is displayed in a block to the lower right.

CUT SCORE CALCULATION TOOL														
Course/Certification Name:			FAIRNESS RESEARCH 3			Test Name:			TEST NAME					
Facilitator Name/Date:			Facilitator Name/Date:			Revision 1 Facilitator Name/Date:			Date: mm/dd/yyyy					
Enter Topic/TPO/Subject ID:			Topic 1			Revision 2 Facilitator Name/Date:								
This spreadsheet tool is the intellectual property of and Copyright ©2019-2020 by Compass Consultants, LLC. Use is limited to the terms of the End User License Agreement (EULA). This copy is limited to 30 DAY DEMO ONLY.														
Test Item QID	Enter * If New, R If Revised	Difficulty Metatag	Average Percentage Correct (Angoff Rating)	Expert 1 Name	Expert 2 Name	Expert 3 Name	Expert 4 Name	Expert 5 Name	Expert 6 Name	Expert 7 Name	Expert 8 Name	Expert 9 Name	Expert 10 Name	Standard Deviation
1.0 M1		Moderate	48.57	60	60	40	45	40	55	40				9.45
1.0 M2		Moderate	62.14	55	70	60	75	65	60	50				8.59
1.0 M3		Moderate	57.14	50	60	70	60	60	60	40				8.55
	R			50	65	60	65	60	60	40				
1.0 M4		Moderate	62.86	70	70	70	60	60	60	50				7.58
1.0 M5		Moderate	60.00	60	70	70	50	50	70	50				10.00
1.0 E1			77.14	70	85	80	75	70	80	70				8.09
1.0 E2			75.00	70	80	90	75	70	80	60				8.57
1.0 E3			77.14	70	80	90	75	75	80	70				8.99
1.0 E4			84.29	75	90	95	90	75	90	75				8.88
1.0 M6		Moderate	62.14	55	70	50	50	70	70	70				8.94
1.0 H1		Hard	41.43	30	50	35	35	50	50	40				8.52
1.0 M7		Moderate	65.57	60	65	70	50	65	65	70				6.90
1.0 M8		Moderate	55.00	55	60	65	50	50	65	40				9.13
	R			70	80	75	80	70	90	40				
1.0 M9		Moderate	57.14	50	50	70	50	50	70	60				8.51
1.0 M10		Moderate	48.57	50	45	50	35	55	55	50				8.90
1.0 M11		Moderate	57.86	50	70	50	50	65	60	60				8.09



Topic Cut Score 58.00 Moderate Difficulty

Approximate Difficulty Rating
 25 - 48.5 : Hard
 48.4 - 71.7 : Moderate
 71.8 - 95 : Easy

Standard Deviation
 A standard deviation of more than 10 will trigger an alert. Discuss the outliers with the judges who set them to determine why. Change as necessary.

7	Easy	In this section	17%
28	Moderate	In this section	67%
7	Hard	In this section	17%
42	TOTAL		100%

Table A- 2 - Cut Score Calculation Tool Data Entry and Display

As data is entered, descriptive statistics are calculated for each judges' ratings as illustrated by table A-3.

SECTION 1 DESCRIPTIVE STATISTICS							
Judge	Number of Items Counted	Minimum Rating	Maximum Rating	Judge's Mean Rating	Judge's Standard Error of the Mean	Judge's Min/Max SD	SD of Judge From Cut Score
1	39	25	75	55.67	0.32	35.36	1.99
2	39	35	90	60.56	0.23	38.89	1.46
3	39	35	95	61.33	0.32	42.43	2.01
4	39	25	90	53.89	0.52	45.96	3.25
5	39	30	85	59.11	0.07	38.89	0.44
6	39	40	90	64.78	0.71	35.36	4.45
7	39	25	75	53.67	0.55	35.36	3.41
8	39	0	0			0.00	
9	39	0	0			0.00	
10	39	0	0			0.00	
Calculated Cut Score (Mean)		58.49	Standard Error of the Mean for Entire Section		1.38		
Average Standard Deviation for Unit		8.64	The standard error of the mean, also called the standard deviation of the mean, is used to estimate the standard deviation of a sampling distribution. The smaller the error, the more reliable the measurement.				
Number of Items Retired (not counted)		3					

Table A- 3 - Judges' Descriptive Statistics

The descriptive statistics are copied for each topic/section to a separate worksheet "CONSOLIDATED DESCRIPTIVE STATISTICS" that displays all of the descriptive statistics for the entire workbook on one worksheet (Figure A-4 – partial view).

SECTION 1 DESCRIPTIVE STATISTICS							
Judge	Number of Items Counted	Minimum Rating	Maximum Rating	Judge's Mean Rating	Judge's Standard Error of the Mean	Judge's Min/Max SD	SD of Judge From Cut Score
1	39	25	75	55.67	0.32	35.36	1.99
2	39	35	90	60.56	0.23	38.89	1.46
3	39	35	95	61.33	0.32	42.43	2.01
4	39	25	90	53.89	0.52	45.96	3.25
5	39	30	85	59.11	0.07	38.89	0.44
6	39	40	90	64.78	0.71	35.36	4.45
7	39	25	75	53.67	0.55	35.36	3.41
8	39	0	0			0.00	
9	39	0	0			0.00	
10	39	0	0			0.00	
Calculated Cut Score (Mean)		58.49	Standard Error of the Mean for Entire Section		1.38		
Average Standard Deviation for Unit		8.64	The standard error of the mean, also called the standard deviation of the mean, is used to estimate the standard deviation of a sampling distribution. The smaller the error, the more reliable the measurement.				
Number of Items Retired (not counted)		3					

SECTION 2 DESCRIPTIVE STATISTICS							
Judge	Number of Items Counted	Minimum Rating	Maximum Rating	Judge's Mean Rating	Judge's Standard Error of the Mean	Judge's Min/Max SD	SD of Judge From Cut Score
1	49	30	75	59.36	0.97	31.82	6.76
2	49	45	90	69.55	0.06	31.82	0.44
3	49	45	95	79.64	0.88	35.36	4.74
4	49	35	95	64.82	0.42	42.43	2.91
5	49	50	80	69.36	0.04	21.11	0.31
6	49	55	90	76.36	0.75	34.75	5.26
7	49	40	80	65.91	0.31	18.28	2.14
8	49	0	0			0.00	
9	49	0	0			0.00	
10	49	0	0			0.00	
Calculated Cut Score (Mean)		68.93	Standard Error of the Mean for Entire Section		1.23		
Average Standard Deviation for Unit		8.64	The standard error of the mean, also called the standard deviation of the mean, is used to estimate the standard deviation of a sampling distribution. The smaller the error, the more reliable the measurement.				
Number of Items Retired (not counted)		9					

Table A- 4 - Consolidated Descriptive Statistics Sample

In addition to the consolidated view of the descriptive statistics, a table comparing the judges' ratings for the entire database is presented (Table A-5). This table is useful to determine if individual judges have typically scored higher or lower than the rest of the group and whether consideration should be given to disregard that judge's scores.

JUDGES COMPARITIVE RATINGS			
Judge	Judge's Mean Rating For All Sections	Judge's Standard Deviation From Assessment Cut Score	Judge's Standard Error of the Mean
1	57.94	4.29	0.37
2	65.24	0.87	0.08
3	70.98	4.94	0.43
4	59.69	3.05	0.26
5	63.11	0.63	0.05
6	71.72	5.46	0.47
7	60.87	2.21	0.19
8			
9			
10			
NOTE: These calculations assume that the judge has contributed input to all topics being analyzed. If the judge did not contribute, their average rating will not be included in this overall rating table - refer to the individual section descriptive statistics for their information.			

Table A- 5 - Judges' Comparative Ratings

The results of each or the worksheets are consolidated on the "FINAL TEST DESIGN" worksheet (Table A-6) which displays totals and percentages of items at each difficulty level for each topic as well as presents a recommended test design to maintain the projected cut score as well as ensure every iteration of the assessment that is generated in a stratified-random format is equal in content and content difficulty. Each column is explained below unless it is self-explanatory.


Final Directed-Randomized Test Design Blueprint for: TEST NAME																mm/dd/yyyy	
Topic	Topic Cut Score & Difficulty	Items in Topic	% of Total Items	Available Hard	% From Topic	Available Mod	% From Topic	Available Easy	% From Topic	Total # Needed From Topic	Use Hard (Calculated)	Use Hard (Actual)	Use Mod (Calculated)	Use Mod (Actual)	Use Easy (Calculated)	Use Easy (Actual)	Topic
Topic 1	58	42	31.34%	7	17%	28	67%	7	17%	6.27	1.04	1	4.18	4	1.04	1	Topic 1
Topic 2	69	52	38.81%	1	2%	27	52%	24	46%	7.76	0.15	1	4.03	4	3.58	3	Topic 2
Topic 3	66	40	29.85%	2	5%	24	60%	14	35%	5.97	0.30	1	3.58	3	2.09	2	Topic 3
4.1		0	0.00%	0		0		0		0.00							4.1
5.1		0	0.00%	0		0		0		0.00							5.1
6.1		0	0.00%	0		0		0		0.00							6.1
7.1		0	0.00%	0		0		0		0.00							7.1
8.1		0	0.00%	0		0		0		0.00							8.1
9.1		0	0.00%	0		0		0		0.00							9.1
10.1		0	0.00%	0		0		0		0.00							10.1
11.1		0	0.00%	0		0		0		0.00							11.1
12.1		0	0.00%	0		0		0		0.00							12.1
13.1		0	0.00%	0		0		0		0.00							13.1
14.1		0	0.00%	0		0		0		0.00							14.1
15.1		0	0.00%	0		0		0		0.00							15.1
16.1		0	0.00%	0		0		0		0.00							16.1
17.1		0	0.00%	0		0		0		0.00							17.1
18.1		0	0.00%	0		0		0		0.00							18.1
19.1		0	0.00%	0		0		0		0.00							19.1
20.1		0	0.00%	0		0		0		0.00							20.1
TOTAL		134	100.00%	10		79		45		20.00	1.49	3	11.79	11	6.72	6	
										NOTE: If appears in the "Total # Needed From Topic" block - you do not have sufficient items in the topic indicated to design a fair test.							
		Test Difficulty Moderate	Test Cut Score 64.00			Set Desired Test Size 20	After all cut-score session data has been entered on section worksheets, set the desired test size in the block to the left. Based upon the number of available items, the quantity of Hard, Moderate and Easy from each section will populate automatically. Use these results to design the test in your test item database using established difficulty Metatags or sub-topic Approximate Difficulty Ratings. Note: Due to rounding errors in Excel, the unit/item difficulty totals may require you to round up or down manually to achieve desired test size. Set the actual number desired based upon the calculated results in the columns labeled "Actual" above. The Checksum to the left will alert you if the selected value does not match the desired test size.										
		Approximate Test Time in Minutes Based on Item Difficulty	17.43			CheckSum 20											

Table A- 6 - Final Test Design Display

- Column D – displays the percentage of items from each topic in relation to the total items available.
- Columns E through J – display the total number and percentage of items at each level of difficulty in each topic. **NOTE:** Percentages are rounded for display.
- Column K – presents the total number of items required from each topic to maintain the percentages shown in column D for the desired test size entered in the “**Set Desired Test Size**” block at the bottom of column F.
- Columns L, N, and P – presents the recommended number of items from each topic at each level of difficulty to be used to maintain the desired cut score as well as topic coverage. These numbers were not rounded to allow the test designer to make informed decisions as to how many items to enter in columns M, O, and Q.

- Columns M, O, and Q – The test designer enters the whole number of items as close to the recommended numbers as possible while referring to the “**Checksum**” block below the desired test size block at the bottom of column F. If there are more items selected than the test size, the checksum block will alert the designer by changing color (Figure A2). The designer must then make an informed decision as to which topic(s) to subtract or add items to ensure topic equity.

Set Desired Test Size
20
Checksum
22

Figure A- 1 - Test Design Sheet Checksum Warning

- The note at the bottom of columns J through R explains a warning that will appear in the “**Total needed from topic**” block(s) if there is not a sufficient quantity of items available to generate a fair test of the size desired.
- The box labeled “Approximate Test Time in Minutes Based on Item Difficulty” is calculated using rule of thumb values for the time it takes a test-taker to respond to a test item based upon its difficulty (see table 3).

Note: The spreadsheet tool is available for licensing. Send request to: info@gocompassconsultants.com

LIST OF FIGURES IN APPENDIX A

Figure A- 1 - Test Design Sheet Checksum Warning	A-7
--	-----

LIST OF TABLES IN APPENDIX A

Table A- 1 - Difficulty Values Based on Number of Alternatives	A-1
Table A- 2 - Cut Score Calculation Tool Data Entry and Display	A-3
Table A- 3 - Judges' Descriptive Statistics.....	A-4
Table A- 4 - Consolidated Descriptive Statistics Sample	A-4
Table A- 5 - Judges' Comparative Ratings	A-5
Table A- 6 - Final Test Design Display	A-6

REFERENCES

- AERA; APA; NCME. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association (AERA); American Psychological Association (APA); National Council on Measurement in Education (NCME). Washington, DC: American Educational Research Association.
- Cizek, G. J. (2006). Standard Setting. (S. Downing, & T. Haladyna, Eds.) *Handbook of test development*, 225 - 258.
- Coscarelli, W., Barrett, A., Kleeman, J., & Shrock, S. (2005). The problem of saltatory cut-score: some issues and recommendations for. *Proceedings of the 9th CAA Conference*. Loughborough: Loughborough University. Retrieved from <https://hdl.handle.net/2134/1984>.
- Dictionary.com Unabridged. (2020, March 6). *empirical*. Retrieved March 6, 2020, from Dictionary.com: <https://www.dictionary.com/browse/empirical>
- Downing, S. M. (2006). Twelve Steps for Effective Test Development. (S. M. Downing, & T. M. Haladyna, Eds.) *Handbook of Test Development*, 3-25.
- Great Schools Partnership. (2013, August 29). *Measurement Error*. Retrieved from The Glossary of Education Reform for Journalists, Parents, and Community Members: <https://www.edglossary.org/measurement-error/>
- Higgins, P. (2009, May). *Item Difficulty and Time Usage*. Retrieved March 2020, from Measurement Research Associates, Inc.: <https://www.rasch.org/mra/mra-05-09.htm>
- Livingston, S. A., & Zieky, M. J. (1982). *Passing scores - A manual for setting standards on educational and occupational tests*. Princeton, NJ: Educational Testing Service.
- Shrock, S. A., & Coscarelli, W. C. (2007). *Criterion-Referenced Test Development - Technical and Legal Guidelines for Corporate Training* (3 ed.). San Francisco, CA, USA: Pfeiffer.
- United States Coast Guard. (2015). *Training system standard operation procedure 10 - Testing (SOP-10)*. Washington, DC, US: United States Coast Guard Forces Command.
- Vale, C. D. (2006). Computerized Item Banking. (S. Downing, & T. Haladyna, Eds.) *Handbook of test development*, 261 - 285.

ACKNOWLEDGEMENTS

The author gratefully acknowledges:

John Kleeman for his friendship, review and recommendations.

John Kleeman, is Executive Director and Founder of Questionmark. John has a first-class degree from Trinity College, Cambridge, and is a Chartered Engineer. John wrote the first version of the Questionmark assessment software system and then founded Questionmark in 1988 to market, develop and support it. John has been heavily involved in assessment software development for over 20 years and has also participated in several standards initiatives: he was on the original team that created IMS QTI and was the instigator and chairman of the panel that produced the Standard BS 7988, which has now become ISO 23988.

Eric Shepherd for his friendship, review, encouragement and publishing guidance.

Eric Shepherd has led international businesses and associations focused on talent, assessments, and success. Eric currently leads Talent Transformation Guild which is focused on helping leaders, talent management professionals, and consultants manage the changes that are now being driven by the pandemic and the technology driven-revolutions. The Guild uses the vocabulary of the Talent Transformation Pyramid to help members understand, learn, and take action, to prepare for the new world of work. Eric stepped away from a CEO role in which he built a SaaS company into a multi-million-dollar international assessment software business. Eric has also led industry and standards initiatives to promote best practices for competency development, assessments, learning, and interoperability.

Sandra Parry, my loving wife, for putting up with my long hours spent tucked away in my office working on this paper.

ABOUT THE AUTHOR

Jim Parry, a 22+ year veteran of the United States Coast Guard, is the Owner and Chief Executive Manager of Compass Consultants, LLC with over 38 years' experience in course design, development, presentation and assessment design and analysis. He holds a Master of Education degree from the University of West Florida and is a Certified Performance Technologist (CPT), awarded by the International Society of Performance Improvement (ISPI). Jim has been a presenter of pre-conference workshops and educational sessions at various professional conferences for many years. He is an internationally recognized consultant providing services concerning test design, development, establishment of cut scores, and analysis. Some of Jim's recent major accomplishments include the development of a comprehensive standard operating procedure (SOP) manual on testing for the U.S. Coast Guard, a spreadsheet tool to assist in recording and analyzing the results of test-centered cut score setting meetings and using the spreadsheet tool to recommend the design of assessments using directed randomization in order to maintain fairness. Jim is a consulting partner of Questionmark Corporation.